

# Complete Specification and Design for Simile

Anthony Hunt

January 3, 2026

## Contents

<b>1</b>	<b>Overview</b>	<b>3</b>
1.1	High Level Ideas . . . . .	3
<b>2</b>	<b>Motivation</b>	<b>3</b>
<b>3</b>	<b>Language</b>	<b>3</b>
3.1	Lexicon . . . . .	4
3.1.1	Whitespace . . . . .	4
3.1.2	Identifiers . . . . .	4
3.1.3	Primitives . . . . .	4
3.1.4	Expressions . . . . .	4
3.1.5	Reserved Keywords . . . . .	6
3.2	Syntax . . . . .	7
3.3	Static Analysis - Type System . . . . .	10
3.3.1	Base Environment Interactions . . . . .	10
3.3.2	Subtype Interactions . . . . .	10
3.3.3	The Set Type . . . . .	11
3.3.4	Primitive Types . . . . .	11
3.3.5	Collection Types . . . . .	11
3.3.6	Relation Subtypes . . . . .	12
3.3.7	Simple Commands . . . . .	13
3.3.8	Quantification Expressions . . . . .	14
3.3.9	Boolean Operations . . . . .	14
3.3.10	Equality and Membership . . . . .	15
3.3.11	Set Operations . . . . .	15
3.3.12	Bag Operations . . . . .	16
3.3.13	Relation Operations . . . . .	16
3.3.14	Relational Subtypes After Operations . . . . .	16
3.3.15	Sequence Operations . . . . .	17
3.3.16	Numerical Operations . . . . .	17
3.3.17	Record Types . . . . .	18
3.3.18	Compound Commands . . . . .	18
3.3.19	Built-in Functions . . . . .	19

3.4	Definedness . . . . .	20
3.5	Language Examples . . . . .	20
<b>4</b>	<b>Features and Behaviours</b>	<b>20</b>
<b>5</b>	<b>User-Facing Specification</b>	<b>20</b>
5.1	(Abstract) Data Types . . . . .	20
5.1.1	Primitives . . . . .	20
5.1.2	Sets . . . . .	21
5.1.3	Relations . . . . .	22
5.1.4	Bags . . . . .	23
5.1.5	Sequences . . . . .	24
5.2	Data Type Operations . . . . .	24
5.2.1	Primitives . . . . .	24
5.2.2	Sets . . . . .	24
5.2.3	Relations . . . . .	24
5.2.4	Bags . . . . .	24
5.2.5	Sequences . . . . .	24
<b>6</b>	<b>Implementation Specification</b>	<b>24</b>
6.1	Abstract Data Types . . . . .	24
6.2	Atomic Operations . . . . .	24
6.3	Optimization Frontend . . . . .	24
6.4	Optimization Backends . . . . .	24
<b>7</b>	<b>AST Lowering and Optimization Rules</b>	<b>24</b>
7.1	Syntactic Sugar for Bags . . . . .	24
7.2	Syntactic Sugar for Sequences . . . . .	25
7.3	Builtin Functions . . . . .	25
7.4	Comprehension Construction . . . . .	25
7.5	Disjunctive Normal Form . . . . .	26
7.6	Or-wrapping . . . . .	27
7.7	Generator Selection . . . . .	27
7.8	GSP to Loop . . . . .	28
7.9	Relational Subtyping Loop Simplification . . . . .	29
7.10	Loop Code Generation . . . . .	30
7.11	Replace and Simplify . . . . .	31

# 1 Overview

This document serves as a living specification of the underlying term rewriting system used in the compiler for a modelling-focused programming language. It includes an overview of the language Simile, a user-facing reference, and implementation plans.

## 1.1 High Level Ideas

**General Strategy** A basic strategy to optimize set and relational expressions is:

1. Normalize the expression as a set comprehensions
2. Simplify and reorganize conjuncts of the set comprehension body

**Intuition** The TRS for this language primarily involves lowering collection data type expressions into pointwise boolean quantifications. Breaking down each operation into set builder notation enables a few key actions:

- Quantifications over sets ( $\{x \cdot G \mid P\}$ ) are naturally separated into generators ( $G$ ) and (non-generating) predicates ( $P$ ). For sets, at least one membership operator per top-level conjunction in  $G$  will serve as a concrete element generator in generated code. Then, top level disjunctions will select one membership operation to act as a generator, relegating all others to the predicate level. For example, if the rewrite system observes an intersection of the form  $\{x \cdot x \in S \wedge x \in T\}$ , the set construction operation must iterate over at least one of  $S$  and  $T$ . Then, the other will act as a condition to check every iteration (becoming  $\{x \cdot x \in S \mid x \in T\}$ ).
- By definition of generators in quantification notation, operations in  $G$  must be statements of the form  $x \in S$ , where  $x$  is used in the “element” portion of the set construction. Statements like  $x \notin T$  or checking a property  $p(x)$  must act like conditions since they do not produce any iterable elements.
- Any boolean expression for conditions may be rewritten as a combination of  $\neg$ ,  $\vee$ , and  $\wedge$  expressions. Therefore, by converting all set notation down into boolean notation and then generating code based on set constructor booleans, we can accommodate any form of predicate function.

## 2 Motivation

## 3 Language

This section denotes the design and behaviour of Simile.

Much of the language design is inherited from Event-B and the mathematical language definition in [1].

## 3.1 Lexicon

Since Simile makes extensive use of Event-B-like set notation and otherwise non-ASCII characters, this section contains a list of such characters, translations to ASCII symbols where suitable, and other useful lexing information.

### 3.1.1 Whitespace

Similar to Python, indentation and newline tokens are significant in Simile. In particular, nested code blocks, as in if-statements, for-loops, and procedure definitions, require the use of indentations to signify the beginning and end. Indentation/newlines may also be used within collection-type comprehensions or enumerations, similar to the style seen in JSON. Comments are single-lined and start with a pound symbol (#).

### 3.1.2 Identifiers

Identifiers in Simile follow the standard regular expression, using only ASCII letters, underscores, and numbers:  $[a-zA-Z\_][a-zA-Z\_0-9]^*$ .

### 3.1.3 Primitives

Integers and floating point numbers follow a similar style to modern programming languages. The dash character (–) may be used in either subtraction or as a unary negation. Superscript numbers may be used for constant exponentiation. Booleans (True/False) follow Python style, with capitalized first letters. Strings are delimited with double-quotes (") and may be multilined.

### 3.1.4 Expressions

ASCII tokens will be provided as an alternative to unicode format where suitable.

The following are symbols used for expressions:

Token	ASCII Token	Name
·	.	Center Dot (Quantifier Separator) <sup>1</sup>
.		Dot
,		Comma
:		Colon
;		Semicolon
		Vertical Bar
λ	lambda	Lambda
(		Left Parenthesis
)		Right Parenthesis
<	<<	Left Bracket

---

<sup>1</sup> Anywhere · is valid, . is valid.

Token	ASCII Token	Name
	[	Left Bracket (Alternative)
)	>>	Right Bracket
	]	Right Bracket (Alternative)
{		Left Brace
}		Right Brace
[[	[[	Left Double Bracket
]]	]]	Right Double Bracket
=		Equals
≠	!=	Not Equals
¬	not	Not
	!	Not (Alternative)
^	and	And
∨	or	Or
⇒	=>	Implies
≡	==	Equivalent
≢	!=	Not Equivalent
∀	forall	For All
∃	exists	There Exists
	+	Plus
	-	Minus
	*	Multiplication
	/	Division
	div	Integer Division
	mod	Modulo
	^	Exponent
<		Less Than
≤	<=	Less Than or Equal
>		Greater Than
≥	>=	Greater Than or Equal
∈	in	In (Membership)
∉	not in	Not In (Membership)
∪	\/	Union
∩	/\	Intersection
\	\	Backslash (Set Minus)
×	><	Cartesian Product
⊂	<<:	Subset
⊆	<:	Subset or Equal
⊃	:>>	Superset
⊇	:>	Superset or Equal
⊄	!<<:	Not Subset
⊈	!<:	Not Subset or Equal
⊅	!:>>	Not Superset
⊉	!:>	Not Superset or Equal
∪	union	General Union
∩	intersection	General Intersection

Token	ASCII Token	Name
$\mapsto$	->	Maplet
$\oplus$	<+>	Relation Overriding
$\circ$	circ	Composition
$\text{++}$	++	Concatenation
$-1$	~	Inverse
$\triangleleft$	<	Domain Restriction
$\triangleleft\triangleleft$	<<	Domain Subtraction
$\triangleright$	>	Range Restriction
$\triangleright\triangleright$	>>	Range Subtraction
$\leftrightarrow$	<->	Relation
$\Leftrightarrow$	<<->	Total Relation <sup>2</sup>
$\Rrightarrow$	<->>	Surjective Relation <sup>3</sup>
$\Leftrightarrow\Rightarrow$	<<->>	Total Surjective Relation <sup>4</sup>
$\rightarrow$	+->	Partial Function
$\twoheadrightarrow$	-->	Total Function
$\hookrightarrow$	>+>	Partial Injection
$\twoheadrightarrow$	>->	Total Injection
$\twoheadrightarrow$	+->>	Partial Surjection
$\twoheadrightarrow$	-->>	Total Surjection
$\twoheadrightarrow$	>->>	Bijection
$\dots$		Upto
$\mathbb{P}$	powerset	Powerset
$\mathbb{P}_1$	powerset1	Non-Empty Powerset

The following are symbols used for programming constructs:

(Unicode) Token	(ASCII) Token	Name
$\coloneqq$	$\coloneqq$	Assignment
$\in$	$::$	Choice (Non-deterministic Membership) Assignment
$\rightarrow$	$\rightarrow$	Right Arrow

### 3.1.5 Reserved Keywords

The following list of symbols and identifiers are reserved for programming constructs:

<sup>2</sup>Brave Sans Mono unicode: U+E100

<sup>3</sup>Brave Sans Mono unicode: U+E101

<sup>4</sup>Brave Sans Mono unicode: U+E102

true	false	if	else	for
while	record	enum	procedure	is
is not	return	break	continue	
from	import	skip	with	type
min	max	choice		

### 3.2 Syntax

We follow the standard EBNF notation for describing the language structure.

Start ::= Statements

Statements ::= {SimpleStmt | CompoundStmt}

SimpleStmt ::= Expr [[ ':' Expr] ( '=' | '∈' ) Expr] NEWLINE [WithStmt]  
 | ControlFlowStmt NEWLINE  
 | ImportStmt NEWLINE

WithStmt ::= INDENT 'with' Expr NEWLINE {Expr NEWLINE} DEDENT

ControlFlowStmt ::= 'return' [Expr]  
 | 'break'  
 | 'continue'  
 | 'skip'

ImportStmt ::= 'import' STRING  
 | 'from' STRING 'import' ImportList

ImportList ::= '\*'  
 | FlatTupleIdentifier

FlatTupleIdentifier ::= IDENTIFIER { ',' IDENTIFIER }  
 | '(' IDENTIFIER { ',' IDENTIFIER } ')'

Expr ::= Quantification | Predicate

Quantification ::= 'λ' [FlatTupleIdentifier] '·' Predicate '|' Expr  
 | '∪' QuantificationBody  
 | '∩' QuantificationBody

QuantificationBody ::= IdentList '·' Predicate '|' Expr  
 | Expr '|' Predicate

IdentList ::= IdentListItem { ',' IdentListItem }

IdentListItem ::= IDENTIFIER  
 | IdentListItem '↦' IdentListItem  
 | '(' IdentList ')'

```

Predicate ::= UnquantifiedPredicate
|   '∀' IdentList '.' Predicate
|   '∃' IdentList '.' Predicate

UnquantifiedPredicate ::= Implication
|   UnquantifiedPredicate ('≡' | '≠') Implication

Implication ::= Disjunction
|   Disjunction {'⇒' Disjunction}

Disjunction ::= Conjunction {'∨' Conjunction}

Conjunction ::= Negation {'∧' Negation}

Negation ::= AtomBool
|   '¬' Negation

AtomBool ::= PairExpr [( '=' | '≠' | '<' | '>' | '≤' | '≥' | '∈' | '∉' | '⊂' | '⊆' |
    '⊃' | '⊇' | '⊄' | '⊈' | '⊅' | '⊉' ) PairExpr]

PairExpr ::= RelSetExpr
|   PairExpr '↦' RelSetExpr

RelSetExpr ::= SetExpr [( '↔' | '↔↔' | '↔↔' | '↔↔' | '↔' | '→' | '↗' | '↘' | '↠'
    | '↠' | '↠' ) RelSetExpr]

SetExpr ::= IntervalExpr
|   IntervalExpr {'∪' IntervalExpr}
|   IntervalExpr {'×' IntervalExpr}
|   IntervalExpr {'⊕' IntervalExpr}
|   IntervalExpr {'∘' IntervalExpr}
|   IntervalExpr {'∩' IntervalExpr}
|   IntervalExpr {'++' IntervalExpr}
|   IntervalExpr ('⊲' | '⊳') IntervalExpr {'∩' IntervalExpr} [RelSubExpr]

RelSubExpr ::= ('\ ' | '▷' | '⊳') IntervalExpr

IntervalExpr ::= ArithmeticExpr ['..' ArithmeticExpr]

ArithmeticExpr ::= Term
|   ArithmeticExpr ('+' | '-') Term

Term ::= Factor
|   Term ('*' | '/' | 'div' | 'mod') Factor

Factor ::= [( '+' | '-' )] Power

```



$\text{Power} ::= \text{Primary} [\text{'^'} \text{Factor} \mid [\text{'-'}](\text{'INTEGER'} \mid \text{'FLOAT'})]$

$\text{Primary} ::= \text{Atom}$   
 $\mid \text{Atom ',' IDENTIFIER}$   
 $\mid \text{Atom '(' Expr \{' , ' Expr \}' ')'}$   
 $\mid \text{Atom '[' Expr ']'}$

$\text{Atom} ::= \text{INTEGER} \mid \text{FLOAT} \mid \text{STRING} \mid \text{BOOLEAN} \mid \text{IDENTIFIER}$   
 $\mid \text{Set} \mid \text{Sequence} \mid \text{Bag} \mid \text{Tuple}$   
 $\mid \text{'P'} \text{'(' Expr ')'} \mid \text{'P}_1 \text{'(' Expr ')'} \mid \text{'(' Expr ')'}$

$\text{Set} ::= \text{'{' CollectionBody '}'}$

$\text{Bag} ::= \text{'[[' CollectionBody ']]'}$

$\text{Sequence} ::= \text{'<' CollectionBody '>'}$

$\text{Tuple} ::= \text{'(' ')'} \mid \text{'(' [NEWLINE INDENT] Expr \{' , ' [NEWLINE [INDENT]]$   
 $\text{[EnumerationBody] ')}'$

$\text{CollectionBody} ::= \text{QuantificationBody} \mid \text{EnumerationBody}$

$\text{EnumerationBody} ::= \text{[NEWLINE INDENT] Expr \{' , ' [NEWLINE [INDENT]]$   
 $\text{Expr} \} \text{[NEWLINE DEDENT]}$

$\text{CompoundStmt} ::= \text{IfStmt} \mid \text{ForStmt} \mid \text{WhileStmt}$   
 $\mid \text{RecordStmt} \mid \text{ProcedureStmt}$

$\text{IfStmt} ::= \text{'if' Predicate ':' Block [ElseStmt]}$

$\text{ElseStmt} ::= \text{'else :'} \text{Block}$   
 $\mid \text{'else if' Predicate ':' Block [ElseStmt]}$

$\text{ForStmt} ::= \text{'for' IdentList '∈' Expr ':' Block}$

$\text{WhileStmt} ::= \text{'while' Expr ':' Block}$

$\text{RecordStmt} ::= \text{'record' IDENTIFIER ':' NEWLINE INDENT (}$   
 $\text{'skip'} \mid \text{TypedName \{' ( , ' [NEWLINE] \mid NEWLINE) TypedName \}}$   
 $\text{) NEWLINE DEDENT}$

$\text{ProcedureStmt} ::= \text{'procedure' IDENTIFIER '(' TypedName \{' , ' TypedName$   
 $\text{\}' ') '→' Expr ':' Block}$

$\text{TypedName} ::= \text{IDENTIFIER [':' Expr]}$

$\text{Block} ::= \text{SimpleStmt}$   
 $\mid \text{NEWLINE INDENT Statements DEDENT}$

### 3.3 Static Analysis - Type System

Let  $\Gamma$  be the type environment set consisting of elements either *identifier* : *Type* or *Type*. The typing rules have been organized in distinct categories for readability. The notation  $Type[x, y, z]$  denotes refinements on *Type* either through instantiating parametric (generic) types or restricting values by way of **with** clauses.

Type refinements may be applicable to types with certain traits (ex., min/-max values for ordered types, definedness of expressions with certain values, set sizes).

#### 3.3.1 Base Environment Interactions

**Includes:** Mechanisms for adding/extracting variables from the type environment.

$$\begin{array}{c}
 \overline{\emptyset \vdash \diamond} \quad (\text{Env } \emptyset) \\
 \frac{\Gamma \vdash A \quad I \notin \text{dom}(\Gamma)}{\Gamma, I : A \vdash \diamond} \quad (\text{Env } I) \\
 \frac{\Gamma_1, I : A, \Gamma_2 \vdash \diamond}{\Gamma_1, I : A, \Gamma_2 \vdash I : A} \quad (\text{Fetch Identifier})
 \end{array}$$

#### 3.3.2 Subtype Interactions

**Includes:** Generic subtyping rules.

$$\begin{array}{c}
 \frac{\Gamma \vdash A}{\Gamma \vdash A \leq A} \quad (\text{Reflexive Subtype}) \\
 \frac{\Gamma \vdash A \leq B \quad \Gamma \vdash B \leq C}{\Gamma \vdash A \leq C} \quad (\text{Transitive Subtype}) \\
 \frac{\Gamma \vdash a : A \quad \Gamma \vdash A \leq B}{\Gamma \vdash a : B} \quad (\text{Subsumption}) \\
 \frac{\Gamma \vdash \diamond}{\Gamma \vdash \text{Any}} \quad (\text{Top Type}) \\
 \frac{\Gamma \vdash A}{\Gamma \vdash A \leq \text{Any}} \quad (\text{Sub Top Type}) \\
 \frac{\Gamma \vdash A' \leq A \quad \Gamma \vdash B \leq B'}{\Gamma \vdash A \rightarrow B \leq A' \rightarrow B'} \quad (\text{Sub Function}) \\
 \frac{\Gamma \vdash A \leq B}{\Gamma \vdash \text{set}[A] \leq \text{set}[B]} \quad (\text{Sub Set}) \\
 \frac{\Gamma \vdash A_1 \leq B_1 \quad \Gamma \vdash A_2 \leq B_2}{\Gamma \vdash A_1 \times A_2 \leq B_1 \times B_2} \quad (\text{Sub Product})
 \end{array}$$

$$\frac{\Gamma \vdash A_1 \leq B_1 \dots A_n \leq B_n \quad \Gamma \vdash A_{n+1} \dots A_{n+m}}{\Gamma \vdash \text{Record}[I_1 : A_1, \dots, I_n : A_n, I_{n+1} : A_{n+1}, \dots, I_{n+m} : A_{n+m}] \leq \text{Record}[I_1 : B_1, \dots, I_n : B_n] \quad (\text{Sub Record})}$$

$$\frac{\Gamma \vdash T}{\Gamma \vdash T[\text{with } x] \leq T} \quad (\text{Type Refinement})$$

### 3.3.3 The Set Type

**Includes:** Universal type; everything is a set. The empty set is a part of every type (aside from primitives).

$$\frac{\Gamma \vdash T}{\Gamma \vdash \text{set}[T]} \quad (\text{Powerset})$$

$$\frac{\Gamma \vdash \diamond}{\Gamma \vdash \text{EmptySet}} \quad (\text{Emptyset})$$

$$\frac{\Gamma \vdash \text{EmptySet}, \text{set}[T]}{\text{EmptySet} \leq \text{set}[T]} \quad (\text{Emptyset Bottom})$$

$$\frac{\Gamma \vdash e_1, \dots, e_n : A}{\Gamma \vdash \{e_1, \dots, e_n\} : \text{set}[A]} \quad (\text{Set Enumeration})$$

### 3.3.4 Primitive Types

**Includes:** Basic types built in to the language, like numbers (int and float), strings, and booleans. Although these could all be represented as sets, it is useful to distinguish these common types from other collections or user-defined types.

$$\frac{\Gamma \vdash \diamond}{\Gamma \vdash \text{bool}} \quad (\text{Primitives - bool})$$

$$\frac{\Gamma \vdash \diamond}{\Gamma \vdash \text{int}} \quad (\text{Primitives - int})$$

$$\frac{\Gamma \vdash \diamond}{\Gamma \vdash \text{float}} \quad (\text{Primitives - float})$$

$$\frac{\Gamma \vdash \diamond}{\Gamma \vdash \text{str}} \quad (\text{Primitives - str})$$

$$\frac{\Gamma \vdash \text{int}}{\Gamma \vdash \text{nat} \leq \text{int}[\text{with } \text{min} = 0]} \quad (\text{Nat from int})$$

$$\frac{\Gamma \vdash \text{int}, \text{float}}{\text{int} \leq \text{float}} \quad (\text{Int from float})$$

### 3.3.5 Collection Types

**Includes:** Sequences, relations, enums (frozen sets), and bags (multisets); Syntactic sugar for abstract data types defined in terms of sets.

$$\begin{array}{c}
\frac{\Gamma \vdash T_1, \dots, T_n}{\Gamma \vdash \text{tuple}[T_1, \dots, T_n] \quad \text{tuple}[T_1, \dots, T_n] = T_1 \times \dots \times T_n} \quad (\text{Tuples}) \\
\frac{\Gamma \vdash \diamond}{\Gamma \vdash () : \text{EmptyTuple}} \quad (\text{Empty Tuple}) \\
\frac{\Gamma \vdash () : \text{EmptyTuple}}{\Gamma \vdash \text{EmptyTuple}, T} \quad (\text{Empty Tuple Bottom}) \\
\frac{\Gamma \vdash \text{EmptyTuple} \leq \text{tuple}[T]}{\Gamma \vdash e_1 : A_1, \dots, e_n : A_n} \quad (\text{Tuple Enumeration}) \\
\frac{\Gamma \vdash e_1 : A_1, \dots, e_n : A_n}{\Gamma \vdash (e_1, \dots, e_n) : \text{tuple}[A_1, \dots, A_n]} \quad (\text{Tuple Enumeration}) \\
\frac{\Gamma \vdash \text{set}[T_1], \text{set}[T_2]}{\Gamma \vdash T_1 \leftrightarrow T_2 \quad T_1 \leftrightarrow T_2 = \text{set}[\text{tuple}[T_1, T_2]]} \quad (\text{Relation Type}) \\
\frac{\Gamma \vdash T, \text{nat}}{\Gamma \vdash \text{bag}[T] \leq T \leftrightarrow \text{nat}} \quad (\text{Bag Type}) \\
\frac{\Gamma \vdash e_1, \dots, e_n : A}{\Gamma \vdash \llbracket e_1, \dots, e_n \rrbracket : \text{bag}[A]} \quad (\text{Bag Enumeration}) \\
\frac{\Gamma \vdash T, \text{nat}}{\Gamma \vdash \text{sequence}[T] \leq \text{nat} \leftrightarrow T} \quad (\text{Sequence Type}) \\
\frac{\Gamma \vdash e_1, \dots, e_n : A}{\Gamma \vdash \langle e_1, \dots, e_n \rangle : \text{sequence}[A]} \quad (\text{Sequence Enumeration}) \\
\frac{\Gamma \vdash s_1, \dots, s_n : \text{str} \quad A = \{s_1, \dots, s_n\} \quad \text{immutable}(A)}{\Gamma \vdash \text{enum}[A] \leq \text{set}[A, \text{with } \text{immutable}]} \quad (\text{Enum from static set})
\end{array}$$

### 3.3.6 Relation Subtypes

**Includes:** Relation subtypes examine four key properties of relations: totality ( $t$ ), surjectivity ( $t_r$ ), one-to-manyness ( $1-$ ), and many-to-oneness ( $1-r$ ). Syntactic sugar to produce refined relations with these properties follows the definitions presented in Event-B. See Section 5.1.3 for more details.

$$\begin{array}{c}
\frac{\Gamma \vdash \text{set}[T_1], \text{set}[T_2]}{\Gamma \vdash T_1 \leftrightarrow T_2 \quad T_1 \leftrightarrow T_2 = \text{set}[\text{tuple}[T_1, T_2]]} \quad (\text{Relation Subtype - Total Relation}) \\
\frac{\Gamma \vdash \text{set}[T_1], \text{set}[T_2]}{\Gamma \vdash T_1 \leftrightarrow T_2 \quad T_1 \leftrightarrow T_2 = \text{set}[\text{tuple}[T_1, T_2]]} \quad (\text{Relation Subtype - Surjective Relation}) \\
\frac{\Gamma \vdash \text{set}[T_1], \text{set}[T_2]}{\Gamma \vdash T_1 \leftrightarrow T_2 \quad T_1 \leftrightarrow T_2 = \text{set}[\text{tuple}[T_1, T_2]]} \quad (\text{Relation Subtype - Total Surjective Relation}) \\
\frac{\Gamma \vdash \text{set}[T_1], \text{set}[T_2]}{\Gamma \vdash T_1 \rightarrow T_2 \quad T_1 \rightarrow T_2 = \text{set}[\text{tuple}[T_1, T_2]]} \quad (\text{Relation Subtype - Partial Function})
\end{array}$$

$$\begin{array}{c}
\frac{\Gamma \vdash \text{set}[T_1], \text{set}[T_2]}{\Gamma \vdash T_1 \rightarrow T_2 \quad T_1 \rightarrow T_2 = (T_1 \leftrightarrow T_2)[\text{with } t, \text{with } 1\bar{-}]} \quad (\text{Relation Subtype - Total Function}) \\
\frac{\Gamma \vdash \text{set}[T_1], \text{set}[T_2]}{\Gamma \vdash T_1 \mapsto T_2 \quad T_1 \mapsto T_2 = (T_1 \leftrightarrow T_2)[\text{with } t, \text{with } 1\bar{-}]} \quad (\text{Relation Subtype - Partial Injection}) \\
\frac{\Gamma \vdash \text{set}[T_1], \text{set}[T_2]}{\Gamma \vdash T_1 \hookrightarrow T_2 \quad T_1 \hookrightarrow T_2 = (T_1 \leftrightarrow T_2)[\text{with } t, \text{with } 1\bar{-}]} \quad (\text{Relation Subtype - Total Injection}) \\
\frac{\Gamma \vdash \text{set}[T_1], \text{set}[T_2]}{\Gamma \vdash T_1 \twoheadrightarrow T_2 \quad T_1 \twoheadrightarrow T_2 = (T_1 \leftrightarrow T_2)[\text{with } t, \text{with } 1\bar{-}]} \quad (\text{Relation Subtype - Partial Surjection}) \\
\frac{\Gamma \vdash \text{set}[T_1], \text{set}[T_2]}{\Gamma \vdash T_1 \rightarrowtail T_2 \quad T_1 \rightarrowtail T_2 = (T_1 \leftrightarrow T_2)[\text{with } t, \text{with } 1\bar{-}]} \quad (\text{Relation Subtype - Total Surjection}) \\
\frac{\Gamma \vdash \text{set}[T_1], \text{set}[T_2]}{\Gamma \vdash T_1 \xrightarrow{\sim} T_2 \quad T_1 \xrightarrow{\sim} T_2 = (T_1 \leftrightarrow T_2)[\text{with } t, \text{with } 1\bar{-}]} \quad (\text{Relation Subtype - Bijection})
\end{array}$$

### 3.3.7 Simple Commands

**Includes:** Assignment, type assignment, type refinements, simple programming language commands/statements.

$$\begin{array}{c}
\frac{\Gamma \vdash E : A \quad I \notin \text{dom}(\Gamma) \vee \Gamma \vdash I : A}{\Gamma \vdash I : A := E} \quad (\text{Variable Assignment}) \\
\frac{\Gamma \vdash T \quad I \notin \text{dom}(\Gamma)}{\Gamma \vdash I := T} \quad (\text{Type Alias Assignment}) \\
\frac{\Gamma \vdash I := T}{\Gamma, I \vdash \diamond \quad I = T} \quad (\text{Type Alias}) \\
\frac{\Gamma \vdash E : A, E_1 : A_1, \dots, E_n : A_n \quad (I \notin \text{dom}(\Gamma) \vee \Gamma \vdash I : A)}{\Gamma \vdash I : A[\text{with } E_1 : A_1, \dots, \text{with } E_n : A_n]} \quad (\text{Refined Variable Assignment}) \\
\frac{\Gamma \vdash T, E_1 : A_1, \dots, E_n : A_n \quad I \notin \text{dom}(\Gamma)}{\Gamma \vdash I := T[\text{with } E_1 : A_1, \dots, \text{with } E_n : A_n]} \quad (\text{Refined Type Alias}) \\
\frac{\Gamma \vdash \diamond}{\Gamma \vdash \text{break} : C} \quad (\text{Command - break}) \\
\frac{\Gamma \vdash \diamond}{\Gamma \vdash \text{continue} : C} \quad (\text{Command - continue}) \\
\frac{\Gamma \vdash \diamond}{\Gamma \vdash \text{skip} : C} \quad (\text{Command - skip}) \\
\frac{\Gamma \vdash E : A}{\Gamma \vdash (\text{return } E) : C} \quad (\text{Command - return})
\end{array}$$

### 3.3.8 Quantification Expressions

**Includes:** Set, sequence, relation, and bag comprehensions, in addition to (set based) lambda expressions. Comprehension variables may be bound through either an *Identifier* or an arbitrarily nested tuple of *Identifiers*

$$\begin{array}{c}
\frac{\Gamma \vdash E : A, P : \text{bool} \quad I_1 : T_1, \dots, I_n : T_n \text{ are vars in } E}{\Gamma \vdash (\lambda I_1, \dots, I_n \cdot P \mid E) : (T_1, \dots, T_n) \rightarrow A} \quad (\text{Lambda Expression}) \\
\\
\frac{\begin{array}{c} \Gamma \vdash E : A, P : \text{bool} \\ I_1, \dots, I_n : \text{TupleIdentifier} \\ Ids = \llbracket identifiers(I_i) \cdot 0 \leq i \leq n \rrbracket \\ \forall e \mapsto count \in Ids \cdot count = 1 \wedge e \notin dom(\Gamma) \\ M = \{I_i \mapsto T \cdot 0 \leq i \leq n \\ \quad \wedge \exists g \cdot \Gamma \vdash g : set[T] \\ \quad \wedge structuralMatch(e, T) \quad (\text{Quantification Body } ^5) \\ \quad \wedge hasGenerator(P, e \in g) \\ \quad \wedge structuralMatchInAllORs(e, T, P) \} \end{array}}{\Gamma \vdash (I_1, \dots, I_n \cdot P \mid E) : QuantificationBody[M, A]} \\
\frac{\Gamma \vdash (I \cdot P \mid E) : QuantificationBody[M, A]}{\Gamma \vdash (\bigcup I \cdot P \mid E) : A} \quad (\text{General Union}) \\
\frac{\Gamma \vdash (I \cdot P \mid E) : QuantificationBody[M, A]}{\Gamma \vdash (\bigcap I \cdot P \mid E) : A} \quad (\text{General Intersection}) \\
\frac{\Gamma \vdash (I \cdot P) : QuantificationBody[M, bool]}{\Gamma \vdash (\forall I \cdot P) : bool} \quad (\text{Forall}) \\
\frac{\Gamma \vdash (I \cdot P) : QuantificationBody[M, bool]}{\Gamma \vdash (\exists I \cdot P) : bool} \quad (\text{Exists}) \\
\frac{\Gamma \vdash (I \cdot P \mid E) : QuantificationBody[M, A]}{\Gamma \vdash \{I \cdot P \mid E\} : set[A]} \quad (\text{Set Comprehension}) \\
\frac{\Gamma \vdash (I \cdot P \mid E) : QuantificationBody[M, A]}{\Gamma \vdash \llbracket I \cdot P \mid E \rrbracket : bag[A]} \quad (\text{Bag Comprehension}) \\
\frac{\Gamma \vdash (I \cdot P \mid E) : QuantificationBody[M, A]}{\Gamma \vdash [I \cdot P \mid E] : sequence[A]} \quad (\text{Sequence Comprehension})
\end{array}$$

### 3.3.9 Boolean Operations

**Includes:** Common predicate logic operations (equivalence, implication, and, or, etc.).

Let  $\otimes = \{\equiv, \neq, \implies, \wedge, \vee\}$

$$\frac{\Gamma \vdash b_1, b_2 : \text{bool}}{\Gamma \vdash b_1 \otimes b_2 : \text{bool}} \quad (\text{Binary Boolean Operations})$$

$$\frac{\Gamma \vdash b_1 : bool}{\Gamma \vdash \neg b_1 : bool} \quad (\text{Boolean Operations - Negation})$$

### 3.3.10 Equality and Membership

**Includes:** Equality and ordering comparisons, set membership.

Let  $\otimes = \{=, \neq\}$

$$\frac{\Gamma \vdash a_1 : T_1, a_2 : T_2}{\Gamma \vdash a_1 \otimes a_2 : bool} \quad (\text{Equals})$$

Let  $\otimes = \{<, \leq, >, \geq\}$

$$\frac{\Gamma \vdash a_1, a_2 : T \quad T \in \{int, float, nat\}}{\Gamma \vdash a_1 \otimes a_2 : bool} \quad (\text{Ordering Operators})$$

Let  $\otimes = \{\in, \notin\}$

$$\frac{\Gamma \vdash a_1 : T_1, a_2 : set[T_2]}{\Gamma \vdash a_1 \otimes a_2 : bool} \quad (\text{Set Membership})$$

### 3.3.11 Set Operations

**Includes:** Subset comparisons, union, intersection, cartesian product, etc.  
Also includes Maplet and Range.

Let  $\otimes = \{\subset, \subseteq, \supset, \supseteq, \not\subset, \not\subseteq, \not\supset, \not\supseteq\}$

$$\frac{\Gamma \vdash a_1 : set[T_1], a_2 : set[T_2]}{\Gamma \vdash a_1 \otimes a_2 : bool} \quad (\text{Set Operations})$$

Let  $\otimes = \{\cup, \cap, \setminus\}$

$$\frac{\Gamma \vdash a_1, a_2 : set[T]}{\Gamma \vdash a_1 \otimes a_2 : set[T]} \quad (\text{Set Operations})$$

$$\frac{\Gamma \vdash a_1 : set[T_1], a_2 : set[T_2]}{\Gamma \vdash a_1 \times a_2 : T_1 \leftrightarrow T_2} \quad (\text{Cartesian Product})$$

$$\frac{\Gamma \vdash a_1 : T_1, a_2 : T_2}{\Gamma \vdash a_1 \mapsto a_2 : tuple[T_1, T_2]} \quad (\text{Maplet})$$

$$\frac{\Gamma \vdash a_1, a_2 : int}{\Gamma \vdash a_1 .. a_2 : set[int, \text{ with } min = a_1, \text{ with } max = a_2]} \quad (\text{Numerical Range})$$

$$\frac{\Gamma \vdash a : T}{\Gamma \vdash \mathbb{P}(a) : set[T]} \quad (\text{Set Operations - Powerset})$$

### 3.3.12 Bag Operations

**Includes:** Bag union, bag intersection, etc.

Let  $\otimes = \{\cup, \cap, +, -\}$

$$\frac{\Gamma \vdash a_1, a_2 : \text{bag}[T]}{\Gamma \vdash a_1 \otimes a_2 : \text{bag}[T]} \quad (\text{Bag Operations - (Max) Union})$$

$$\frac{\Gamma \vdash a : T_1 \leftrightarrow T_2, b : \text{bag}[T_1]}{\Gamma \vdash a[b] : \text{bag}[T_2] \quad a[b] = a^{-1} \circ b} \quad (\text{Bag Operations - Image})$$

### 3.3.13 Relation Operations

**Includes:** Relation overriding, composition, domain/range manipulation. See Section 5.1.3 for changes in relational subtypes after applying operations.

$$\frac{\Gamma \vdash a : T_1 \leftrightarrow T_2, b : T_1}{\Gamma \vdash a(b) : T_2} \quad (\text{Relation Operations - Function Call})$$

$$\frac{\Gamma \vdash a : T_1 \leftrightarrow T_2, b : \text{set}[T_1]}{\Gamma \vdash a[b] : \text{set}[T_2]} \quad (\text{Relation Operations - Image})$$

$$\frac{\Gamma \vdash a_1, a_2 : T_1 \leftrightarrow T_2}{\Gamma \vdash a_1 \oplus a_2 : T_1 \leftrightarrow T_2} \quad (\text{Relation Operations - Overriding})$$

$$\frac{\Gamma \vdash a_1 : T_1 \leftrightarrow T_2, a_2 : T_2 \leftrightarrow T_3}{\Gamma \vdash a_1 \circ a_2 : T_1 \leftrightarrow T_3} \quad (\text{Relation Operations - Composition})$$

$$\frac{\Gamma \vdash a_1 : T_1, a_2 : T_1 \leftrightarrow T_2}{\Gamma \vdash a_1 \triangleleft a_2 : T_1 \leftrightarrow T_2} \quad (\text{Relation Operations - Domain Restriction})$$

$$\frac{\Gamma \vdash a_1 : T_1, a_2 : T_1 \leftrightarrow T_2}{\Gamma \vdash a_1 \trianglelefteq a_2 : T_1 \leftrightarrow T_2} \quad (\text{Relation Operations - Domain Subtraction})$$

$$\frac{\Gamma \vdash a_1 : T_1 \leftrightarrow T_2, a_2 : T_2}{\Gamma \vdash a_1 \triangleright a_2 : T_1 \leftrightarrow T_2} \quad (\text{Relation Operations - Range Restriction})$$

$$\frac{\Gamma \vdash a_1 : T_1 \leftrightarrow T_2, a_2 : T_2}{\Gamma \vdash a_1 \trianglerighteq a_2 : T_1 \leftrightarrow T_2} \quad (\text{Relation Operations - Range Subtraction})$$

### 3.3.14 Relational Subtypes After Operations

**Includes:** Changes in relational subtyping (totality, surjectivity, etc.) after applying operations. For simplicity, we implement subtype changes in a few binary operators as a boolean mask representing: [with total ( $t$ ), with total on range ( $t_r$ ), with one-to-manyness ( $1-$ ), with many-to-oneness ( $1-r$ )]. Subtyping is done conservatively: we keep subtype properties only when they are guaranteed. See Section 5.1.3 for more on relational subtyping.

$$\frac{\Gamma \vdash s : \text{set}[T_1], r : (T_1 \leftrightarrow T_2)[\text{True}, x, y, z]}{\Gamma \vdash (s \triangleleft r) : (T_1 \leftrightarrow T_2)[\text{False}, x, y, z]} \quad (\text{Relational Subtype - Domain Restriction})$$



$$\begin{array}{c}
\frac{\Gamma \vdash s : \text{set}[T_1], r : (T_1 \leftrightarrow T_2)[\text{True}, x, y, z]}{\Gamma \vdash (s \triangleleft r) : (T_1 \leftrightarrow T_2)[\text{False}, x, y, z]} \text{ (Relational Subtype - Domain Subtraction)} \\
\frac{\Gamma \vdash s : \text{set}[T_2], r : (T_1 \leftrightarrow T_2)[x, \text{True}, y, z]}{\Gamma \vdash (r \triangleright s) : (T_1 \leftrightarrow T_2)[x, \text{False}, y, z]} \text{ (Relational Subtype - Range Restriction)} \\
\frac{\Gamma \vdash s : \text{set}[T_2], r : (T_1 \leftrightarrow T_2)[x, \text{True}, y, z]}{\Gamma \vdash (r \triangleright s) : (T_1 \leftrightarrow T_2)[x, \text{False}, y, z]} \text{ (Relational Subtype - Range Subtraction)} \\
\frac{\Gamma \vdash r : (T_1 \leftrightarrow T_2)[w, x, y, z]}{\Gamma \vdash r^{-1} : (T_1 \leftrightarrow T_2)[x, w, z, y]} \text{ (Relational Subtype - Inverse)} \\
\frac{\Gamma \vdash r : (T_1 \leftrightarrow T_2)[a, b, c, d], t : (T_1 \leftrightarrow T_2)[w, x, y, z]}{\Gamma \vdash (r \oplus t) : (T_1 \leftrightarrow T_2)[w, a \oplus b, c \wedge d, x]} \text{ (Relational Subtype - Overriding)} \\
\frac{\Gamma \vdash r : (T_1 \leftrightarrow T_2)[a, b, c, d], t : (T_2 \leftrightarrow T_3)[w, x, y, z]}{\Gamma \vdash (r \circ t) : (T_1 \leftrightarrow T_3)[a, b \wedge c, c \wedge y, d \wedge z]} \text{ (Relational Subtype - Composition)}
\end{array}$$

### 3.3.15 Sequence Operations

**Includes:** Concatenation

$$\frac{\Gamma \vdash a_1, a_2 : \text{sequence}[T]}{\Gamma \vdash a_1 \mathbin{++} a_2 : \text{sequence}[T]} \text{ (Sequence Operations - Concatenation)}$$

### 3.3.16 Numerical Operations

**Includes:** Addition, subtraction, multiplication, etc. on natural numbers, integers, and floats.

$$\begin{array}{c}
\frac{\Gamma \vdash a_1 : T_1, a_2 : T_2 \quad T_i \in \{\text{int}, \text{nat}\}}{\Gamma \vdash a_1 \text{ div } a_2 : \max(T_1, T_2)} \text{ (Integer Operations - Division)} \\
\frac{\Gamma \vdash a_1 : T_1, a_2 : T_2 \quad T_i \in \{\text{int}, \text{nat}\}}{\Gamma \vdash a_1 \text{ mod } a_2 : \max(T_1, T_2)} \text{ (Integer Operations - Modulo)} \\
\frac{\Gamma \vdash a_1 : T_1, a_2 : T_2 \quad T_i \in \{\text{int}, \text{nat}, \text{float}\}}{\Gamma \vdash a_1 + a_2 : \max(T_1, T_2)} \text{ (Numerical Operations - Addition)} \\
\frac{\Gamma \vdash a_1 : T_1, a_2 : T_2 \quad T_i \in \{\text{int}, \text{nat}, \text{float}\}}{\Gamma \vdash a_1 - a_2 : \max(T_1, T_2, \text{int})} \text{ (Numerical Operations - Subtraction)} \\
\frac{\Gamma \vdash a_1 : T_1, a_2 : T_2 \quad T_i \in \{\text{int}, \text{nat}, \text{float}\}}{\Gamma \vdash a_1 * a_2 : \text{float}} \text{ (Numerical Operations - Floating Division)} \\
\frac{\Gamma \vdash a_1 : T_1, a_2 : T_2 \quad T_i \in \{\text{int}, \text{nat}, \text{float}\}}{\Gamma \vdash a_1 \text{ div } a_2 : \max(T_1, T_2)} \text{ (Numerical Operations - Multiplication)}
\end{array}$$

$$\begin{array}{c}
\frac{\Gamma \vdash a_1 : T \quad T \in \{int, nat, float\}}{\Gamma \vdash -a_1 : max(int, T)} \quad (\text{Numerical Operations - Negation}) \\
\frac{\Gamma \vdash a_1 : T_1, a_2 : T_2 \quad T_i \in \{int, nat, float\}}{\Gamma \vdash a_1 \wedge a_2 : max(T_1, T_2)} \quad (\text{Numerical Operations - Exponentiation})
\end{array}$$

### 3.3.17 Record Types

**Includes:** Types for named tuples.

$$\begin{array}{c}
\frac{\Gamma \vdash a : Record[I : A, \dots]}{\Gamma \vdash a.I : A} \quad (\text{Records - Access}) \\
\frac{\Gamma \vdash A_1, \dots, A_n \quad \forall I_i, I_j \cdot I_i \neq I_j}{\Gamma \vdash Record[I_1 : A_1, \dots, I_n : A_n]} \quad (\text{Records - Type Definition}) \\
\frac{\Gamma \vdash a_1 : A_1, \dots, a_n : A_n}{\Gamma \vdash record(a_1 = A_1, \dots, a_n = A_n) : Record[a_1 : A_1, \dots, a_n : A_n]} \quad (\text{Records - Initialization})
\end{array}$$

### 3.3.18 Compound Commands

**Includes:** If, While, For, and Import statements, along with Procedure definitions and calls.

$$\begin{array}{c}
\frac{\Gamma \vdash C_1, C_2}{\Gamma \vdash C_1 \setminus n C_2} \quad (\text{Command - Composition}) \\
\frac{\Gamma \vdash \diamond}{\Gamma \vdash \text{import } M : C} \quad (\text{Command - Valid Import Module}) \\
\frac{\Gamma \vdash \diamond}{\Gamma \vdash \text{from } M \text{ import } I_1, \dots, I_n : C} \quad (\text{Command - Valid Import Names}) \\
\frac{\Gamma \vdash \text{import } M \text{ (with identifiers and types } I_1 : A_1, \dots, I_n : A_n)}{\Gamma, M = record(I_1 = A_1, \dots, I_n = A_n) \vdash \diamond} \quad (\text{Command - Import Module}) \\
\frac{\Gamma \vdash \text{from } M \text{ import } I_1, \dots, I_n \text{ (with types } A_1, \dots, A_n)}{\Gamma, I_1 : A_1, \dots, I_n : A_n \vdash \diamond} \quad (\text{Command - Import Names}) \\
\frac{\Gamma \vdash C, E, A_1, \dots, A_n}{\Gamma \vdash \text{procedure } I = C \therefore I : procedure[A_1, \dots, A_n, E]} \quad (\text{Command - Procedure Definition}) \\
\frac{\Gamma \vdash procedure[A_1, \dots, A_n, E], a_1 : A_1, \dots, a_n : A_n}{\Gamma \vdash I(a_1, \dots, a_n) : E} \quad (\text{Command - Procedure Call}) \\
\frac{\Gamma \vdash C, b : bool}{\Gamma \vdash \text{if } b : C} \quad (\text{Command - If}) \\
\frac{\Gamma \vdash C_1, C_2, b : bool}{\Gamma \vdash \text{if } b : C_1 \text{ else } : C_2} \quad (\text{Command - If Else})
\end{array}$$

$\frac{\Gamma \vdash C, b : \text{bool}}{\Gamma \vdash \text{while } b : C}$	(Command - While)
$\frac{\Gamma \vdash C, p : \text{bool} \quad p = i \in G}{\Gamma \vdash \text{for } p : C}$	(Command - For)
$\frac{\Gamma \vdash D, I : A \quad \Gamma, I : A \vdash C}{\Gamma \vdash C \text{ substitute vars in } D}$	(Command - Block)

### 3.3.19 Built-in Functions

**Includes:** Common set-related functions.

$\frac{\Gamma \vdash T[\text{with } \text{order} \in \text{traits}(T)]}{\Gamma \vdash \text{min} : \text{set}[T] \rightarrow T}$	(Built-in - Minimum)
$\frac{\Gamma \vdash T, \text{int}}{\Gamma \vdash \text{min} : \text{set}[T] \times (T \rightarrow \text{int}) \rightarrow T}$	(Built-in - Mapped Minimum)
$\frac{\Gamma \vdash T[\text{with } \text{order} \in \text{traits}(T)]}{\Gamma \vdash \text{max} : \text{set}[T] \rightarrow T}$	(Built-in - Maximum)
$\frac{\Gamma \vdash T, \text{int}}{\Gamma \vdash \text{max} : \text{set}[T] \times (T \rightarrow \text{int}) \rightarrow T}$	(Built-in - Mapped Maximum)
$\frac{\Gamma \vdash T}{\Gamma \vdash \text{choice} : \text{set}[T] \rightarrow T}$	(Built-in - Choice)
$\frac{\Gamma \vdash T}{\Gamma \vdash \text{dom} : T_1 \leftrightarrow T_2 \rightarrow \text{set}[T_1]}$	(Built-in - Domain)
$\frac{\Gamma \vdash T}{\Gamma \vdash \text{ran} : T_1 \leftrightarrow T_2 \rightarrow \text{set}[T_2]}$	(Built-in - Range)
$\frac{\Gamma \vdash T}{\Gamma \vdash \text{card} : \text{set}[T] \rightarrow \text{nat}}$	(Built-in - Cardinality)
$\frac{\Gamma \vdash T}{\Gamma \vdash \text{size} : \text{bag}[T] \rightarrow \text{nat}}$	(Built-in - Bag Size)
$\frac{\Gamma \vdash T \quad T \in \{\text{int}, \text{float}, \text{nat}\}}{\Gamma \vdash \text{sum} : \text{set}[T] \rightarrow T}$	(Built-in - Sum)
$\frac{\Gamma \vdash T_1, T_2}{\Gamma \vdash \text{cast} : \text{type}[T_1] \times T_1 \rightarrow T_2}$	(Built-in - Cast)
$\frac{\Gamma \vdash T_1, T_2, x : E}{\Gamma \vdash \text{cast} : \text{type}[T_1] \times T_1 \times \text{set}[E] \rightarrow T_2[\text{with } x : E]}$	(Built-in - Cast With <sup>6</sup> )

## 3.4 Static Analysis - Type System Extensions with Traits

### 3.4.1 Trait Collection

Traits are used to augment types, providing the compiler with additional context for the use of an abstract type that could influence the generated implementa-

tion. The following paragraphs list identified traits and provides justification for their existence.

**Constant Pre-computation** The type may contain values for easily identifiable constants, allowing for specialized generated instructions. When possible, constant traits will propagate along operations until the constant values can no longer be guaranteed.

- *Literal*: A trait containing a single literal value. If a type has this trait, it is guaranteed to be of this value.
- *Domain*: A looser form of *Literal* that carries a set of possible values. When applied, the type is restricted to (one of) the values from this *Domain*.
- Leave out for now: *UnknownElementInDomain*: Loosens the *Domain* guarantee so that values in the domain are only likely to appear, but not guaranteed? We would need to see if this opens up any optimizations...
- *Immutable*: Represents types whose values (and value contents) cannot be modified. Should we instead assume immutability and have a *Mutable* trait?

**Arithmetic** Restricts the value of primitive types. Collection types wishing to take advantage of this information should instead store it in the underlying primitive element type. For example, applying *Min* to a *set[int]* should be written as *set[int][with min = 0]*, not *set[int,with min = 0]* (although we can loosen this with syntactic sugar). Like Event-B, *set* types are really powersets of the underlying element type, whereas primitive types represent values chosen from a *set* of primitive values.

- *Orderable*: Values of this type can be ordered. This may be useful for order-based implementations like B-Trees for sets.
- *Min*: minimum value
- *Max*: maximum value

### Collection Traits

- *Iterable*: All collections are iterable (they can be used as a for-loop generator).
- *Empty*: A collection with guaranteed cardinality = 0.
- *Size*: An over-estimate of a set's size. When available, *Size* will never under-estimate the number of elements in a collection.

- *UniqueElements*: All entries are guaranteed to be unique. More useful for the domain/range of relations, checking if a bag only has range = 1, or if a sequence has no repeats.
- *Total*: The elements in a set cover the entirety of the domain of the underlying element type. On relations, this means every possible pair combination is an element.

**Relations** Represents the relational subtype properties (injectiveness, surjectiveness, etc.)

- *TotalOnDomain*: All elements of the domain are within the mapping.
- *TotalOnRange*: All elements of the range are within the mapping.
- *ManyToOne*: Elements of the domain may be repeated with different range values.
- *OneToMany*: Elements of the range may be repeated with different range values. *ManyToOne* and *OneToMany* imply a one-to-one mapping

## Generics

- *GenericBound*: Contains the set of types a generic may become. For example,  $T[\text{with } \textit{GenericBound} = \{int, float\}]$  represents any numeric value.

### 3.4.2 Trait-trait Interactions

Collection of trait-implies-trait interactions that are universal (i.e., they do not depend on a type). Most traits are really just boolean values, as in, does this type have this property. Thus when traits used with boolean operations should be interpreted as truthy/falsy.

Interpret these as if it were type initialization rules. So, if a type were initialized with these traits, we would get the following implications “for-free”. A type can only have one of each trait (except for multiple literals?), so traits given in the assumption can be used in the result without ambiguity or naming. Traits with function notation (parenthesis) are assumed to be new constructions.

TODO design decision: should traits be applied automatically or just verified? If we apply automatically, we could give best effort and then error/remove traits that are incompatible. In verification, we just throw an error. We may use trait restrictions per-type to further determine incompatibilities (ex. an integer cannot be empty?).

$$\begin{array}{c}
 \textit{Domain} = \{\} \\
 \hline
 \neg \textit{Domain} \\
 \textit{Literal} \quad \neg \textit{Domain} \\
 \hline
 \textit{Domain} := \{\textit{Literal}\}
 \end{array}
 \begin{array}{l}
 \text{(Domain Empty)} \\
 \\
 \text{(Literal implies a Domain)}
 \end{array}$$

$\frac{Literal \quad Domain}{Domain := Domain \cup Literal}$	(Literal within Domain)
$\frac{Domain \quad Orderable \quad \neg Min}{Min := min(Domain)}$	(Orderable Domain without Min)
$\frac{Domain \quad Orderable \quad Min}{Min := min(Min, Domain)}$	(Orderable Domain with Min)
$\frac{Domain \quad Orderable \quad \neg Max}{Max = max(Domain)}$	(Orderable Domain without Max)
$\frac{Domain \quad Orderable \quad Max}{Max = max(Max, Domain)}$	(Orderable Domain with Max)
$\frac{Min}{Orderable}$	(Min implies Order)
$\frac{Max}{Orderable}$	(Max implies Order)
$\frac{UniqueElements \quad Size = len(Domain) \quad \neg Empty}{Total}$	(Full set)
$\frac{Size = 0}{Empty}$	(Empty Size)
$\frac{Size \neq 0}{\neg Empty}$	(Non-Empty Size)
$\frac{Size}{Iterable}$	(Size implies Iterable)

### 3.4.3 Type-trait interactions

Some types are incompatible with some traits, and some types must have some traits. This section lists the possible traits for each type - asterisks are for mandatory traits, and traits not listed are not allowed for that specific type.

**All types** Literal and Domain support constant propagations and compile time computations.

- Immutable
- Literals
- Domain

#### Bool

- Domain = {true, false}\*

#### None

- Literal\*
- Domain? (implied from literal)

### **String**

- Empty
- Size (length)
- Unique?
- Iterable\*
- Orderable (lexicographic)\*

### **Int, Float**

- Min
- Max
- Size (num of bits? or just depend on min/max? If so, min/max must be given, with a default value of 64 bit range)
- Iterable\*
- Orderable\*

### **Tuple**

- Min
- Max
- Size (length)
- UniqueElements
- Total
- Iterable\*
- Orderable (lexicographic if dependents are orderable?)

**Maplet** Tuple but with the addition of:

- Domain?
- Size = 2\*

### **Record**

- Iterable\* (treat as a relation)
- Domain\* (str entries)
- ManyToOne\* (treat as a relation with keys = strings)

### **Procedure**

- Domain should be stored in the domain of the input types since there are too many to count? Range should be the domain of the return type
- Need to worry about generics?

### **Set**

- Orderable if children are
- Min
- Max
- Iterable\*
- Empty
- UniqueElements\*
- Size
- Total

**Relation** All traits except generic trait

**Bag** Same as relation, with ManyToOne\*. Should we add a trait called Count, which is the sum of the range values?

**Seq** Same as relation, with ManyToOne\*, Size\*

### **Enum**

1. Immutable\*
2. Domain\*
3. Literal\*
4. Iterable\*
5. UniqueElements\*
6. Size\*?
7. Can never be total unless enum size = 1, then always total



### 3.4.4 Operation-Trait Interactions

Operations on *Literal* traits will always produce a *Literal* where possible, following the effects of applying the function on values. For conciseness, we omit all *Literal* interactions and instead refer to the semantics of the operations themselves.

$$\begin{array}{c}
\frac{a : T[Max = x] \quad b : T[Min = y] \quad x < y}{(a < b) : Bool[Literal = true]} \quad \text{(Less than with max-min)} \\
\frac{a : T[Max = x] \quad b : T[Min = y] \quad x \leq y}{(a \leq b) : Bool[Literal = true]} \quad \text{(Less than or equal with max-min)} \\
\frac{a : T[Min = x] \quad b : T[Max = y] \quad x > y}{(a < b) : Bool[Literal = false]} \quad \text{(Less than with min-max)} \\
\frac{a : T[Min = x] \quad b : T[Max = y] \quad x > y}{(a \leq b) : Bool[Literal = false]} \quad \text{(Less than or equal with min-max)} \\
\frac{a : Int[Literal] \quad b : Int[Literal]}{(a .. b) : set[Int[Min = a, Max = b], Literal = \{a, a + 1, \dots, b\}, Total]} \quad \text{(Upto)} \\
\frac{s : set[Empty]}{(s = \{\}) : Bool[Literal = true]} \quad \text{(Is Empty)} \\
\frac{s : set[T, Size = x] \quad a : T}{(s \cup \{a\}) : set[T, Size = x + 1]} \quad \text{(Set add)} \\
\frac{s : set[T, Size = x, Total] \quad a : T}{(s \cup \{a\}) : set[T, Size = x, Total]} \quad \text{(Set add with Total)} \\
\frac{s : set[T, Size = x] \quad a : T}{(s \setminus \{a\}) : set[T, Size = x]} \quad \text{(Set delete)} \\
\frac{s : set[T, Size = x, Total] \quad a : T}{(s \setminus \{a\}) : set[T, Size = x - 1]} \quad \text{(Set delete removes Total)} \\
\frac{s : set[T[Domain = S]] \quad a : T[Literal = A]}{(a \in s) : Bool[Literal = A \in S]} \quad \text{(Membership within domain)} \\
\frac{s : set[T, Total] \quad a : T}{(a \in s) : Bool[Literal = true]} \quad \text{(Membership within total domain)} \\
\frac{a_1, a_2, \dots, a_n : T}{\{a_1, a_2, \dots, a_n\} : set[T, Size = n]} \quad \text{(Enumeration)} \\
\frac{s : set[T, Size = n]}{card(s) : Int[Max = n]} \quad \text{(Cardinality)} \\
\frac{s : set[T, Size = n, Total]}{powerset(s) : set[set[T], Size = 2^n, Total]} \quad \text{(Powerset)}
\end{array}$$

$$\begin{array}{c}
\frac{s : \text{set}[T, \text{Empty}]}{\text{choice}(s) : \text{TypeError}} \quad (\text{Empty choice fails (statically)}) \\
\\
\frac{s : \text{set}[T[\text{Max} = y, \text{Size} = n]]}{\text{sum}(s) : T[\text{Max} = y \times n]} \quad (\text{Sum bounds}) \\
\\
\frac{s : \text{set}[T[\text{Max} = y, \text{Size} = n]]}{\text{sum}(s) : T[\text{Max} = y^n]} \quad (\text{Product bounds}) \\
\\
\frac{s : \text{set}[T[\text{Min} = y]] \quad a : T[\text{Literal} = y] \quad a \in s}{\text{min}(s) : T[\text{Literal} = y]} \quad (\text{Min looks for Min}) \\
\\
\frac{s : \text{set}[T[\text{Max} = y]] \quad a : T[\text{Literal} = y] \quad a \in s}{\text{max}(s) : T[\text{Literal} = y]} \quad (\text{Max looks for Max}) \\
\\
\frac{s : \text{set}[T, \text{Literal} = a] \quad t : \text{set}[T, \text{Literal} = b]}{(s \cup t) : \text{set}[T, \text{Literal} = a \cup b]} \quad (\text{Union Literal}) \\
\\
\frac{s : \text{set}[T[\text{Domain} = a]] \quad t : \text{set}[T[\text{Domain} = b]]}{(s \cup t) : \text{set}[T[\text{Domain} = a \cup b]]} \quad (\text{Union with widest Domain}) \\
\\
\frac{s : \text{set}[T[\text{Min} = a]] \quad t : \text{set}[T[\text{Min} = b]] \quad a < b}{(s \cup t) : \text{set}[T[\text{Min} = a]]} \quad (\text{Union with lowest Min}) \\
\\
\frac{s : \text{set}[T[\text{Max} = a]] \quad t : \text{set}[T[\text{Max} = b]] \quad a < b}{(s \cup t) : \text{set}[T[\text{Max} = b]]} \quad (\text{Union with highest Max}) \\
\\
\frac{s : \text{set}[T, \text{Empty}] \quad t : \text{set}[T, x] \quad a < b}{(s \cup t) : \text{set}[T, x]} \quad (\text{Union with Empty keeps Traits}) \\
\\
\frac{s : \text{set}[T, \text{Size} = x] \quad t : \text{set}[T, \text{Size} = y] \quad x < y}{(s \cup t) : \text{set}[T, \text{Size} = y]} \quad (\text{Union with widest size}) \\
\\
\frac{s : \text{set}[T] \quad t : \text{set}[T, \text{Total}]}{(s \cup t) : \text{set}[T, \text{Total}]} \quad (\text{Union with Total stays}) \\
\\
\frac{s : \text{set}[T[\text{GenericBound} = x]] \quad t : \text{set}[T[\text{GenericBound} = y]]}{(s \oplus t) : \text{set}[T[\text{GenericBound} = x \cap y]]} \quad (\text{Generics with set ops}) \\
\\
\frac{s : \text{set}[T[\text{Domain} = x]] \quad t : \text{set}[V[\text{Domain} = y]]}{(s \times t) : \text{set}[\text{tuple}[T, V, \text{Domain} = x \times y]]} \quad (\text{Cartesian Product}) \\
\\
\frac{s : \text{set}[T, \text{Total}] \quad t : \text{set}[V, \text{Total}]}{(s \times t) : \text{set}[\text{tuple}[T, V, \text{Total}]]} \quad (\text{Total Cartesian Product}) \\
\\
\frac{s : \text{set}[T, \text{Min} = w, \text{Max} = x] \quad t : \text{set}[T, \text{Min} = y, \text{Max} = z] \quad w > y \quad x < z}{(s \subset t : \text{set}[T, \text{Min} = w, \text{Max} = x]) : \text{Bool}} \quad (\text{Subset with min-max}) \\
\\
\frac{r : \text{relation}[T[\text{Domain} = x], V[\text{Domain} = y]]}{r^{-1} : \text{relation}[V[\text{Domain} = y], T[\text{Domain} = x]]} \quad (\text{Inverse swaps Domain})
\end{array}$$

$$\begin{array}{c}
\frac{s : \text{relation}[T, V, \text{Empty}] \quad t : \text{relation}[V, U]}{s \circ t : \text{relation}[T, U, \text{Empty}]} \quad (\text{Empty Composition}) \\
\frac{r : \text{relation}[T, V] \quad s : \text{set}[T, \text{Empty}]}{r[s] : \text{set}[V, \text{Empty}]} \quad (\text{Image with Empty Arg}) \\
\frac{r : \text{relation}[T, V, \text{Empty}] \quad s : \text{set}[T]}{r[s] : \text{set}[V, \text{Empty}]} \quad (\text{Image with Empty Relation}) \\
\frac{r : \text{relation}[T, V, \text{TotalOnDomain}] \quad s : \text{set}[T, \text{Domain} = \text{domain}(r), \text{Size} > 0]}{r[s] : \text{set}[V, \text{Size} > 0]} \quad (\text{Image with ManyToOne}) \\
\frac{r : \text{relation}[T, V, \text{ManyToOne}] \quad s : \text{set}[T, \text{Size} = n]}{r[s] : \text{set}[V, \text{Size} = n]} \quad (\text{Image with ManyToOne}) \\
\frac{r : \text{relation}[T, V, \text{TotalOnDomain}]}{\text{domain}(r) : \text{set}[T, \text{Total}]} \quad (\text{Total domain}) \\
\frac{r : \text{relation}[T, V, \text{TotalOnRange}]}{\text{range}(r) : \text{set}[V, \text{Total}]} \quad (\text{Total range}) \\
\frac{s : \text{sequence}[T, \text{Size} = x] \quad t : \text{sequence}[T, \text{Size} = y]}{s \mathbin{++} t : \text{sequence}[T, \text{Size} = x + y]} \quad (\text{Concat with Size})
\end{array}$$

### 3.5 Definedness

### 3.6 Language Examples

## 4 Features and Behaviours

## 5 User-Facing Specification

### 5.1 (Abstract) Data Types

For the most part, Simile's semantics resemble those of set theory and specification languages (specifically, Event-B). Aside from a few primitive types that abstract over bare set theory and programming language specific structures, like integer arithmetic and procedure definitions, all objects inherit properties from sets. For example, a sequence can be modelled as a set of pairs from natural numbers to the sequence element type.

In the forthcoming sections, we explicitly state the design of these set-theory-inspired types, useful properties, and underlying implementations.

#### 5.1.1 Primitives

With consideration to execution efficiency and the extensive amount of existing optimization in modern languages, Simile makes use of the following basic types:

**Booleans** As the cornerstone of logical reasoning, booleans deserve an explicit type with reserved identifiers:  $\mathbb{B} = \{True, False\}$ . Further, Simile provides a healthy supply of common operations beyond  $\wedge$  and  $\vee$ :  $\implies, \equiv, \longleftarrow, \forall, \exists$ , etc.

**Integers** While integers are a very mature type in the world of computing, trade-offs between arbitrary-precision and fixed-point arithmetic may require testing and careful review.

If Simile chooses to represent numbers through conventional fixed point arithmetic, struggles with overflow and extremely large numbers are compounded by the innate non-determinism within Simile. For example, calculating the sum of a set of very large numbers can be done in arbitrary order, according to the semantics of mathematics (and Simile). But in fixed-point arithmetic, the resulting value can very well differ based on the order of operations. In other words, commutative math operations are no longer commutative under fixed precision. At some point in the computation pipeline, Simile will need to decide the order in which to evaluate the set, which may cause hard-to-spot bugs and unexpected runtime failures.

Conversely, choosing arbitrary precision integers may bring Simile operators closer to mathematics at the cost of speed. The core features of Simile, while targeting formal methods users, attempt to efficiently execute set and relation operations. Adding checks for overflow on every mathematical operation may prove detrimental to performance.

The implementation decision essentially is one of context vs. safety. To benchmark Simile with these types, we gather a list of examples and judge whether the performance hit is worth the reliability benefits. Specific questions to examine: How often does overflow occur in these examples? Do numbers approach the upper limits of fixed precision often enough to warrant accommodating infrastructure? What is the performance of fixed vs. arbitrary precision calculations.

**Floating Point Numbers** In a similar vein to integers, floats are an imperfect translation of real numbers into efficiently calculable representations. Since floats (or even reals) are not used as often in specifications as integers, we leave the implementation as the conventional 64-bit representation. However, we note that non-determinism on these operations will prove especially tricky, since the result of commutative complements may differ without explicit over/underflow.

**Strings** Strings are syntactic sugar over Sequences of characters, represented internally as a sequence of (fixed precision) integers. Eventually, Simile aims to support built-in string operations as seen in many common programming languages.

### 5.1.2 Sets

Foundational Simile type: an unordered, unique collection of objects.

Sets may be used in two ways:

1. Enumerations by way of  $\{x, y, z\}$  (read: The set of elements  $x, y$ , and  $z$ )
2. Comprehensions of the form  $\{x \cdot x \in S \mid f(x)\}$  (read: The set of function  $f$  applied to all elements in  $S$ ). A small effort is made to accommodate a shorthand form, wherein  $\{x \cdot x \in S \mid x\}$  may be written as  $\{x \cdot x \in S\}$ , but this syntax easily confuses bound/unbound variables and is not recommended for complex comprehension expressions.

Sets, as mathematical objects with only two major restrictions, lend themselves well to a variety of computational implementations. Arrays, ordered arrays, hash maps, bit sets, roaring bit sets, path maps, and tries can all be used to satisfy the uniqueness property and hide an internally ordered representation from users.

No matter the underlying implementation, the optimizer may make use of syntactically-analyzed set sizes to direct the choice of lowered iterator. Further, recording the cardinality of a set at runtime may prove beneficial to the performance of common cardinality-based specification expressions at a very small memory cost. Other properties based on element type, like max/min for numeric types, limits for numeric ranges, maximum sizes, may be useful.

### 5.1.3 Relations

Relations are sets that make use of paired/maplet elements (i.e., they define a relationship between two sets). Like sets, implementations may vary greatly, with the potential to increase efficiency of relation-specific operations.

Like sets, relations may benefit from compile-time computed properties: size of the domain, size of the range, totality, many-to-oneness,  $O(1)$  access to the domain and range, and access to associated domain/range properties.

**Relational Subtypes** The notion of a relation is a broad classification of a set of pairs, where the domain, range, and relationship are not bound by any conditions. However, when generating code based on purely mathematical relations, it can be advantageous to use notions of totality, injectivity, surjectivity, etc. In particular, these named relational subtypes can be described with four properties: total on the domain, total on the range, one-to-many, and many-to-one.

The below table converts conventional notation into these four subtypes:

Relational Subtype	Properties			
	T	TR	1-	1-R
Relation				✓
Partial Function			✓	
Partial Injection			✓	✓
Surjective Relation		✓		
		✓		✓
Partial Surjection		✓	✓	
		✓	✓	✓
Total Relation	✓			
	✓			✓
Total Function	✓		✓	
Total Injection	✓		✓	✓
Total Surjective Relation	✓	✓		
	✓	✓		✓
Total Surjection	✓	✓	✓	
Bijection	✓	✓	✓	✓

Table 3: Conventional mathematical notation of relational subtypes decomposed into four properties: domain totality (T), range totality (TR), domain one-to-manyness (1-), and range one-to-manyness (1-R)

**Subtype Properties** Properties that can affect code generation may involve:

- New properties after applying operations on sets/relations
- Size of result
- Validity of operations
- Deterministic nature of operations
- Super/subset properties
- Relational identities

Below is a list of the impact of these properties.

Domain Totality:

- $R[S] \neq \emptyset$
- $R(S)$  is always valid
- $R^{-1}$  is TR
- $|R| \geq \text{dom}(R)$

Range Totality:

- $R^{-1}$  is T
- $|R| \geq \text{ran}(R)$

Domain One-to-manyness:

- $R(\{e\})$  is deterministic
- $|R[\{e\}]| = 1$
- $|R[S]| \leq |S|$
- $|R| \leq \text{dom}(R)$

Range One-to-manyness:

- $|R| \leq \text{ran}(R)$

#### 5.1.4 Bags

Bags (or Multisets) are a refinement of relations wherein, for a bag with element type  $T$ , the bag is a relational function from  $T \rightarrow \mathbb{Z}^+$ , denoting the count of items in a collection. Once the count of an item reaches 0, the item is removed from the bag.

Bags must be many-to-one and carry specialized behaviours for conventional relation operations. For example, relational image  $(R[B])$  with a bag argument will now return a bag, where the number of occurrences of the resulting item replaces what would otherwise be a regular set of those items.

Additional fields to record the size of the bag (sum of the range) may prove useful.

#### 5.1.5 Sequences

Sequences are yet another refinement of relations where a sequence with element type  $T$  represents a relational function from  $\mathbb{N} \rightarrow T$  that is total on some defined bounds.

The dense nature of sequence domains, along with knowledge of sequence sizes, may open up further data type refinements.

## 5.2 Data Type Operations

### 5.2.1 Primitives

### 5.2.2 Sets

### 5.2.3 Relations

### 5.2.4 Bags

### 5.2.5 Sequences

## 6 Implementation Specification

### 6.1 Abstract Data Types

### 6.2 Atomic Operations

### 6.3 Optimization Frontend

### 6.4 Optimization Backends

## 7 AST Lowering and Optimization Rules

Below is a list of rewrite rules for key abstract data types, syntactic sugar, and some builtin functions. Phases are intended to be executed in order; the post-condition of one phase serves as the pre-condition for the next.

### 7.1 Syntactic Sugar for Bags

$$R[B] \rightsquigarrow R^{-1} \circ B \quad (\text{Bag Image})$$

$$S \cup T \rightsquigarrow \llbracket x \cdot x \in \text{dom}(S) \cup \text{dom}(T) \wedge n = \max(S[x] \cup T[x]) \mid x \mapsto n \rrbracket$$

(Bag Predicates - Union)

$$S \cap T \rightsquigarrow \llbracket x \cdot x \in \text{dom}(S) \cap \text{dom}(T) \wedge n = \min(S[x] \cup T[x]) \wedge n > 0 \mid x \mapsto n \rrbracket$$

(Bag Predicates - Intersection)

$$S + T \rightsquigarrow \llbracket x \cdot x \in \text{dom}(S) \cup \text{dom}(T) \wedge n = \text{sum}(S[x] \cup T[x]) \mid x \mapsto n \rrbracket$$

(Bag Predicates - Sum)

$$S - T \rightsquigarrow \llbracket x \cdot x \in \text{dom}(S) \wedge n = S(x) - (T \cup \llbracket x \mapsto 0 \rrbracket)(x) \wedge n > 0 \mid x \mapsto n \rrbracket$$

(Bag Predicates - Difference)

### 7.2 Syntactic Sugar for Sequences

$$L \mathbin{++} M \rightsquigarrow L \cup \{ x \mapsto y \cdot x \mapsto y \in M \mid x + \max(\text{dom}(L)) \mapsto y \}$$

(Concatenation)

$$\text{flatten}(X) \rightsquigarrow \text{foldL}(\mathbin{++}, [], X) \quad (\text{Flatten})$$

$$\text{foldL}(f, e, X) \rightsquigarrow \text{foldL}(f, f(e, \text{first}(X)), \text{tail}(X)) \quad (\text{Fold Left Associative})$$



$$\begin{aligned}
foldL(f, e, []) &\rightsquigarrow e && \text{(Fold Left Associative - Empty)} \\
first(X) &\rightsquigarrow X(0) && \text{(First)} \\
tail(X) &\rightsquigarrow \{ i \mapsto x \cdot i \mapsto x \in X \wedge i \neq 0 \mid i - 1 \mapsto x \} && \text{(Tail)} \\
append(X, y) &\rightsquigarrow X \cup \{ max(dom(X)) \mapsto y \} && \text{(Append)} \\
prepend(X, y) &\rightsquigarrow \{ i \mapsto x \cdot i \mapsto x \in X \mid i + 1 \mapsto x \} \cup 0 \mapsto y && \text{(Prepend)}
\end{aligned}$$

### 7.3 Builtin Functions

$$\begin{aligned}
f(x) := E &\rightsquigarrow f := f \oplus \{ x \mapsto E \} && \text{(Functional Override)} \\
card(S) &\rightsquigarrow \sum x \cdot x \in S \mid 1 && \text{(Cardinality)} \\
dom(R) &\rightsquigarrow \{ x \mapsto y \cdot x \mapsto y \in R \mid x \} && \text{(Domain)} \\
ran(R) &\rightsquigarrow \{ x \mapsto y \cdot x \mapsto y \in R \mid y \} && \text{(Range)} \\
R \oplus Q &\rightsquigarrow Q \cup (dom(Q) \triangleleft R) && \text{(Override)} \\
S \triangleleft R &\rightsquigarrow \{ x \mapsto y \cdot x \mapsto y \in R \wedge x \in S \mid x \mapsto y \} && \text{(Domain Restriction)} \\
S \triangleleft R &\rightsquigarrow \{ x \mapsto y \cdot x \mapsto y \in R \wedge x \notin S \mid x \mapsto y \} && \text{(Domain Subtraction)} \\
R \triangleright S &\rightsquigarrow \{ x \mapsto y \cdot x \mapsto y \in R \wedge y \in S \mid x \mapsto y \} && \text{(Range Restriction)} \\
R \triangleright S &\rightsquigarrow \{ x \mapsto y \cdot x \mapsto y \in R \wedge y \notin S \mid x \mapsto y \} && \text{(Range Subtraction)} \\
size(S) &\rightsquigarrow \sum x \mapsto n \in S \cdot n && \text{(Bag Size)}
\end{aligned}$$

### 7.4 Comprehension Construction

**Intuition** All set-like variables and literals are decomposed into set comprehensions.

#### Post-condition

- All terms with a set-like type (relations, bags, sets, sequences, etc.) must be in comprehension form.

$$\begin{aligned}
R(x) &\rightsquigarrow choice(R[x]) && \text{(Functional Image } ^a) \\
R[S] &\rightsquigarrow \{ x \mapsto y \cdot x \mapsto y \in R \wedge x \in S \mid y \} && \text{(Image)} \\
x \mapsto y \in S \times T &\rightsquigarrow x \in S \wedge y \in T && \text{(Product)} \\
x \mapsto y \in R^{-1} &\rightsquigarrow y \mapsto x \in R && \text{(Inverse)} \\
x \mapsto y \in (Q \circ R) &\rightsquigarrow x \mapsto z \in Q \wedge z' \mapsto y \in R \wedge z = z' && \text{(Composition)}
\end{aligned}$$

---

<sup>a</sup>May fail if R is not total. *choice* is nondeterministic.

$$\begin{aligned}
S \cup T &\rightsquigarrow \{x \cdot x \in S \vee x \in T\} && \text{(Predicate Operations - Union)} \\
S \cap T &\rightsquigarrow \{x \cdot x \in S \wedge x \in T\} && \text{(Predicate Operations - Intersection)} \\
S \setminus T &\rightsquigarrow \{x \cdot x \in S \wedge x \notin T\} && \text{(Predicate Operations - Difference)} \\
x \in \oplus(E \mid P) &\rightsquigarrow P \wedge x = E && \text{(Membership Collapse <sup>a</sup>)}
\end{aligned}$$

<sup>a</sup>Rule only matches inside the predicate of a quantifier. Explicitly enumerating all matches for all quantification types and predicate cases (ANDs, ORs, etc.) would require too much boilerplate.  $x$  must be bound by the encasing quantifier.

The  $\oplus$  operator represents any quantifier that returns a set-like type (ex. generalized union/intersection, set comprehension, relation comprehension).

## 7.5 Disjunctive Normal Form

**Intuition** All quantifier predicates are expanded to DNF (i.e.  $\wedge$ -operations nested within top-level  $\vee$ -operations).

**Notes** All matches of this phase occur only inside quantifier predicates.

### Post-condition

- All terms with a set-like type (relations, bags, sets, sequences, etc.) must be in comprehension form.
- Quantifier predicates are in disjunctive normal form - top level or-clauses with inner and-clauses.
- If no  $\vee$  operators exist within a quantifier predicate, the predicate must only contain  $\wedge$  operators.

$$x_1 \wedge \dots \wedge (x_i \wedge x_{i+1}) \wedge \dots \rightsquigarrow x_1 \wedge \dots \wedge x_i \wedge x_{i+1} \wedge \dots \quad \text{(Flatten Nested } \wedge \text{)}$$

$$x_1 \vee \dots \vee (x_i \vee x_{i+1}) \vee \dots \rightsquigarrow x_1 \vee \dots \vee x_i \vee x_{i+1} \vee \dots \quad \text{(Flatten Nested } \vee \text{)}$$

$$\neg \neg x \rightsquigarrow x \quad \text{(Double Negation)}$$

$$\neg(x \vee y) \rightsquigarrow \neg x \wedge \neg y \quad \text{(Distribute De Morgan - Or)}$$

$$\neg(x \wedge y) \rightsquigarrow \neg x \vee \neg y \quad \text{(Distribute De Morgan - And)}$$

$$x \wedge (y \vee z) \rightsquigarrow (x \wedge y) \vee (x \wedge z) \quad \text{(Distribute } \wedge \text{ over } \vee \text{)}$$

## 7.6 Or-wrapping

**Intuition** Restructuring quantifier predicates with top-level ORs for later rules.

### Post-condition

- All terms with a set-like type (relations, bags, sets, sequences, etc.) must be in comprehension form.

- Quantifier predicates are in disjunctive normal form - top level or-clauses with inner and-clauses.

$$\{E \cdot \bigwedge P_i\} \rightsquigarrow \{E \cdot \bigvee \bigwedge P_i\} \quad (\text{Or-wrapping } ^a)$$

---

<sup>a</sup>To simplify the matching process later on, we wrap every top-level AND statement (which is guaranteed to be a ListOp by the dataclass field type definition) with an OR.

## 7.7 Generator Selection

**Intuition** All nested ANDs are placed in a tree structure to better suit loop lowering.

### Post-condition

- All terms with a set-like type (relations, bags, sets, sequences, etc.) must be in comprehension form.
- Each top-level or-clause within quantification predicates must have one selected ‘generator’ predicate (GeneratorSelectionPredicate - GSP) of the form  $x \in S$  that loops over its bound dummy variable  $x$ .
- All conjunctive clauses are wrapped in either a GSP or CombGSP structure.
- Quantifier predicates are in disjunctive normal form (with GSP replacing AND structures) - top level or-clauses with inner and-clauses.

**Notes** All matches of this phase occur only inside quantifier predicates. GSP is a tuple of form (generator chain, predicates) that can be flattened to a conjunctive clause. A CombGSP is a triple of form (shared generator, disjunctive children predicates/generators, predicates) that can likewise be flattened into a conjunctive clause. We keep the structure of these tuples distinct to ease the rewriting process.

### Brainstorming selection heuristics

Conditions to consider:

- Prefer composition chains in case of nested loops (ex.  $x \mapsto y \in R \wedge y \mapsto z \in Q$ , but what about renaming? -  $x \mapsto y \in R \wedge y' \mapsto z \in Q \wedge y = y'$ ). Perhaps this could just be a preference for relations over sets if we see a nested quantifier.
- Choose most common generator among a list of or-clauses (allows for greater simplification later on)
- Choose smallest generator (by set size). This information may not always be accessible (ex. if a set is created by reading values from standard input). Functions may need to choose this on a case-by-case basis (ie. one

function call could have two args, with a small set on the left arg and a large set on the right arg. But what if the sizes are reversed later in the code? We've already statically lowered it).

- When should constant folding happen? Equality substitution necessary for this?
- This may need to work with nesting considerations.
- What if we try moving these optimizations to *loop* structures far later in the pipeline?

$$\bigwedge P_i \rightsquigarrow GSP(\bigwedge_{P_i \notin P_g} P_{g_i}, \bigwedge_{P_i \in P_g} P_i) \quad (\text{GSP Wrapping } ^a)$$

$$GSP(x \wedge xs, P) \vee GSP(x \wedge ys, Q) \rightsquigarrow CombGSP(x, GSP(xs, P) \vee GSP(ys, Q)) \quad (\text{Nested Generator Selection } ^b)$$

---

<sup>a</sup>GSP is a tuple-like structure, where all entries can be flattened into an AND. The LH term must occur inside a quantifier's predicate - one generator per top-level or-clause. Chained generators may be necessary to allow for nested loop operations (ie. non-top-level clauses may require generators).  $P_g$  is a list of chained generators, specific set membership clauses (of form  $x \in S$ ) distinguished from the rest of  $\bigwedge P_i$ . Currently, the selection of  $P_g$  is arbitrary (and thus the rewrite system is not confluent), but heuristics may be added later to choose better generators.

<sup>b</sup>*free* is a function that returns true when variables in the predicate are defined (either bound by the current generator or bound prior to the current statement)

## 7.8 GSP to Loop

**Intuition** Start lowering expressions into imperative-like loops.

### Post-condition

- All quantifiers are transformed into *loop* structures.
- *loop* predicates are of the form  $x \in S \wedge \bigwedge P_i$ .
- All *loop* predicates have an assigned generator of the form  $x \in S$ .

$$\oplus E \mid P \rightsquigarrow \begin{array}{l} a := \text{identity}(\oplus) \\ \mathbf{loop} \ P : \\ \quad a := \text{accumulate}(a, E) \end{array} \quad (\text{Quantifier Generation } ^a)$$

---

<sup>a</sup> $\oplus$  works for any quantifier (but not  $\forall$  and  $\exists$ ). The identity and accumulate functions are determined by the realized  $\oplus$ . For example, if  $\oplus = \sum$ , the identity is 0 and accumulate is addition.

$$\begin{array}{ccc}
\text{loop } \bigvee P_i : & \rightsquigarrow & \begin{array}{l} \text{loop } P_0 \\ \text{body} \\ \text{loop } P_1 \wedge \neg P_0 \\ \text{body} \\ \text{loop } P_2 \wedge \neg \bigvee_{i < 2} P_i \\ \text{body} \end{array} \\
\text{body} & & 
\end{array}
\quad \text{(Top-level Or-Loop)}$$

$$\begin{array}{ccc}
\text{loop } GSP(g \wedge gs, P) & \rightsquigarrow & \begin{array}{l} \text{loop} \\ \text{SingleGSP}(g, \text{free}(P)) \\ \text{loop } GSP(gs, \text{bound}(P)) \\ \text{body} \end{array} \\
\text{body} & & 
\end{array}
\quad \text{(Chained GSP Loop } ^a)$$

---

<sup>a</sup> *free* considers defined variables at this point in time.

$$\begin{array}{ccc}
\text{loop } GSP([], P) & \rightsquigarrow & \text{if } P \text{ then} \\
\text{body} & & \text{body} \\
& & \text{(Empty GSP Loop)}
\end{array}$$

$$\begin{array}{ccc}
\text{loop } CombGSP(g, gs, P) & \rightsquigarrow & \begin{array}{l} \text{loop } \text{SingleGSP}(g, P) \\ \text{loop } gs_0 \\ \text{body} \\ \text{loop } gs_1 \\ \text{body} \\ \dots \end{array} \\
\text{body} & & 
\end{array}
\quad \text{(Combined GSP Loop)}$$

## 7.9 Relational Subtyping Loop Simplification

**Intuition** Simplify unnecessary iteration structures into direct accesses.

**Post-condition**

- All quantifiers are transformed into *loop* structures.
- *loop* predicates are of the form  $x \in S \wedge \bigwedge P_i$ .
- All *loop* predicates have an assigned generator of the form  $x \in S$ .

No change from previous, but the number of loop structures should not increase.

$$x \in \text{dom}(R) \rightsquigarrow \text{True} \quad \text{(Total membership elimination } ^a)$$

$$x \in \text{ran}(R) \rightsquigarrow \text{True} \quad \text{(Surjective membership elimination } ^b)$$

---

<sup>a</sup> *R* is total and  $x$  satisfies the type of *R*'s domain. Currently unused since it may eliminate some generators - move higher in the list?

<sup>b</sup> *R* is surjective and  $x$  satisfies the type of *R*'s range. Currently unused since it may eliminate some generators - move higher in the list? Or

$\text{loop } \text{SingleGSP}(x \mapsto y \in R, x = a \wedge P) \quad \rightsquigarrow$	$\text{if } a \in \text{dom}(R) \text{ then}$ $\quad \text{loop } \text{SingleGSP}(y \in R[a], P[x := a])$ $\quad \text{body}$ $\quad \text{(Concrete Domain Image)}$
$\text{loop } \text{SingleGSP}(x \mapsto y \in R, y = a \wedge P) \quad \rightsquigarrow$	$\text{if } a \in \text{ran}(R) \text{ then}$ $\quad \text{loop } \text{SingleGSP}(x \in R^{-1}[a], P[y := a])$ $\quad \text{body}$ $\quad \text{(Concrete Range Image } ^a)$

---

<sup>a</sup>Requires efficient lookups for *ran*, inverse, and image

$\text{loop } \text{SingleGSP}(y \in R[x], P)$ $\quad \text{body}$ <p>where <i>free</i>(<i>x</i>) and <i>R</i> is many-to-one</p>	$\rightsquigarrow$ $\text{if } P[y := R(a)] \text{ then}$ $\quad \text{body}[y := R(a)]$ $\text{(Single Element Loop)}$
$\text{loop } \text{SingleGSP}(x \in \{y\}, P)$ $\quad \text{body}$	$\rightsquigarrow$ $\text{if } P[x := y] \text{ then}$ $\quad \text{body}[x := y]$ $\text{(Singleton Membership Elimination)}$

## 7.10 Loop Code Generation

**Intuition** Eliminate all intermediate loop structures.

**Post-condition**

- AST is in imperative code style (for loops, if statements, etc).
- All *loop* and quantification constructs have been eliminated.
- Some variables may not be defined.

$\text{loop } \text{SingleGSP}(P_g, \bigwedge P_i)$ $\quad \text{body}$	$\rightsquigarrow$ $\text{if } \bigwedge_{\text{free}(P_i)} P_i \text{ then}$ $\quad \text{for } P_g \text{ do}$ $\quad \text{if } \bigwedge_{\text{bound}(P_i)} P_i \text{ then}$ $\quad \quad \text{body}$ $\quad \text{(Conjunct Conditional } ^a)$
---	--

---

<sup>a</sup>Function *free* returns clauses in *P* that contain only free + defined variables. *bound* returns the clauses that contain the bound variable *x* or undefined variables. *P<sub>g</sub>* is the selected generator.

## 7.11 Replace and Simplify

**Intuition** Eliminate all undefined variables (with implicit  $\exists$  quantifiers).

### Post-condition

- AST is in imperative code style (for loops, if statements, etc).
- All variables are defined.

$$\begin{array}{ccc} \text{if } Identifier(x) = E \wedge P \text{ then} & \rightsquigarrow & \text{if } P[x := E] \text{ then} \\ \text{body} & & \text{body}[x := E] \\ & & \text{(Equality Elimination } ^a) \end{array}$$

---

<sup>a</sup> $x$  is an undefined, unbound variable in the current scope.  $E$  is an expression that does not contain  $x$ . Simplify resulting booleans.

$$\begin{array}{ll} P \wedge True \rightsquigarrow P & \text{(Simplify And-true)} \\ P \wedge False \rightsquigarrow False & \text{(Simplify And-false)} \\ P \vee True \rightsquigarrow True & \text{(Simplify Or-true)} \\ P \vee False \rightsquigarrow P & \text{(Simplify Or-false)} \\ Statement(Statement(x)) \rightsquigarrow Statement(x) & \text{(Flatten Nested Statements)} \end{array}$$

### References

- [1] Christophe Métayer and Laurent Voisin. “The Event-B Mathematical Language”. 2007.