

The SKill Language

Timm Felden

TR 2013/06

Abstract

This paper presents an approach to serializing objects, which is tailored for usability, performance and portability. Unlike other general serialization mechanisms, we provide explicit support for extension points in the serialized data, in order to provide a maximum of upward compatibility and extensibility.

Acknowledgements

Main critics: Erhard Plödereder and Martin Wittiger.
Additional critics: Dominik Bruhn¹, Dennis Przytarski².

To do (1)

To do (2)

To do (3)

To do (4)

Contents

I	Specification Language	5
1	Motivation	5
1.1	Scientific Contributions	6
2	Outline	6
3	Related Work	6
4	Syntax	9
4.1	Reserved Words	9
4.2	The Grammar	9
4.3	Examples	10
4.4	Style Guide	13
5	Semantics	14
5.1	A Specification File	14
5.2	Includes	15
5.3	Type Declarations	15
5.4	Field Declarations	16
5.4.1	Constants	16
5.4.2	Transient Fields	17

¹Discussion on SKill and related work

²Discussion on API and proposal of reordering type information; Pragmas

5.4.3	Change-Modifier	17
5.5	Field Types	17
5.5.1	Container Types	18
5.5.2	Annotations	18
5.6	Type Annotations	18
5.7	Name Resolution	18
6	The Type System	19
6.1	Built-In Types	19
6.2	Compound Types	21
6.3	User Types	21
6.3.1	Legal Types	22
6.3.2	Equivalence of Type and Field Names	22
6.4	Fancy Types	22
6.4.1	Interfaces	22
6.4.2	Enums	22
6.5	Views	22
6.6	Typedefs	23
6.7	Default Values	23
6.8	Examples	23
7	Type Annotations	24
7.1	Restrictions	24
7.1.1	Type Restrictions	25
7.1.1.1	Unique	25
7.1.1.2	Singleton	25
7.1.1.3	Monotone	26
7.1.1.4	Abstract	27
7.1.1.5	Default	27
7.1.2	Field Restrictions	27
7.1.2.1	NonNull	27
7.1.2.2	Default	28
7.1.2.3	Range	28
7.1.2.4	Coding	29
7.1.2.5	Constant Length Pointer	30
7.1.2.6	OneOf	30
7.2	Hints	31
7.2.1	Owner	31
7.2.2	Provider	32
7.2.3	Remove Restrictions	32
7.2.4	Constant Mutator	33
7.2.5	Mixin	33
7.2.6	Flat	33
7.2.7	Unique	33
7.2.8	Pure	34
7.2.9	Distributed	34
7.2.10	OnDemand	34
7.2.11	Monotone	35
7.2.12	ReadOnly	35
7.2.13	Ignore	35

7.2.14	Pragma	35
II	User Interface	36
8	In Memory Representation	36
8.1	API	36
8.2	Representation of Objects	36
8.2.1	Proposed Data Structures	36
8.2.2	Reading Unmodified Data	37
8.2.3	Modifying Data	37
8.2.4	Reading Data in a Modified State	38
8.2.5	Writing Data	38
8.2.6	Final Thoughts on Runtime Complexity	38
III	Binary File Format	39
9	Impact of changing Specifications	39
10	Serialization	39
10.1	Steps of the Serialization Process	40
10.2	General File Layout	40
10.2.1	Layout of a String Block	41
10.2.2	Layout of a Type Block	41
10.2.3	Node Example	44
10.3	Storage Pools	45
10.3.1	Subtypes	46
10.3.2	Appending Fields to existing Pools	46
10.3.3	Appending Objects to existing Pools	47
10.4	Serialization of Field Data	47
10.5	Endianness	48
10.6	Date Example	48
10.7	Map Example	50
11	Deserialization	50
11.1	An LL-inspired Approach to sf-Parsing	51
11.1.1	Lexer	51
11.1.2	Parser State	51
11.1.3	Parser	52
11.2	Date Example	53
IV	Future Work	55
V	Appendix	56
A	Full Grammar of the Specification Language	56
B	Full Grammar of the Binary File Format	57

C	Variable Length Coding	58
D	Error Reporting	60
E	Levels of Language Support	61
F	Numerical Limits	62
G	Numerical Constants	62
	Glossary	63
	Acronyms	63

Part I

Specification Language

1 Motivation

To do (5)

To do (6)

To do (7)

Many industrial and scientific projects suffer from platform or language dependent representation of their core data structures. These problems often cause software engineers to stick with outdated tools or even programming languages, thus causing a lot of frustration. This does not only increase the burden of hiring new project members, but can ultimately cause a project to die unnecessarily.

The approach presented in this paper provides means of platform and language independent specification of serializable data structures and therefore a safe way to let old tools of a tool suite talk to the new ones, without even the need of recompiling the old ones. We set out to design a new and easy to use way of making core data structures of a tool platform language independent, because we believe that the best language a programmer can use to write a new tool, is the language that he likes the most. We also had the strict requirement to provide a solution that can describe an intermediate representation with stable parts that can be used for decades and unstable parts that may change on a daily basis.

In order to achieve this goal, we introduce two new concepts:

The first one is an easy to use specification language for data structures providing simple data types like integers and strings, container types like sets and maps, type safe pointers, extension points and single inheritance. The specification language is modular in order to make large specifications more readable.

The second one is a formalized mapping of specified types to a bitwise representation of stored objects. The mapping is very compact and therefore scalable, easy to understand and therefore easy to bind to a new language. It does encode the type system and can therefore provide a maximum of upward and downward compatibility, while maintaining type safety at the same time. It allows for a maximum of safety when it comes to manipulating data unknown to the generated interface, while maintaining high decoding and encoding speeds³.

An improvement over the Extensible Markup Language (XML), our main competitor, is that the reflective usage of stored data is expected to be quite rare, because the Serialization Killer Language (SKill) binding generator is able to generate an interface that ensures type safety of modifications and provides a nice integration into the target language. This leads to a situation, where it is possible to use files containing data of arbitrary types. If the data stored in the file is not used by a client, he does not have to pay for it with execution time or memory. Furthermore a client does not have to know the whole intermediate representation of a tool suite, but only the parts he is going to use in order to achieve his goals. The expected file sizes range from a megabyte to several gigabytes, while having virtually no relevant numerical limits in the file format⁴. Please note, that the SKill file format is a lot more compact than equivalent XML files would be. It is expected, that files contain objects of hundreds

³The serialization and deserialization operations are linear in the size of the input/output file.

⁴There are practical limits, such as Java having array lengths limited to 2^{31} or current file systems having a maximum file size limit that is roughly equivalent to the size of a file completely occupied by objects with a single field of a single byte. There will also be problems with raw I/O-Performance for very large files and an implementation of a binding generator, which can handle files not storable in the main memory is a tricky thing to do.

of types with thousands of instances each. If a type in such a file would contain three pointers on average, the file size would still be around a mega byte, which is due to a very compact representation of stored data. This will also lead to high load and store performance, because the raw disk speed is expected to be the limiting factor.

1.1 Scientific Contributions

This section is a very concise representation of contributions that in part have already been mentioned above and in parts will be mentioned much later.

The suggested serialization format and serialization language offer all of the following features in a single product:

- a small footprint and therefore high decoding speeds
- a fully reflective type encoding
- type safe storage of references both to known and unknown types⁵
- a rich type system providing amongst others references, containers, single inheritance and extension points
- the specification language is modular⁶ and easy to use
- no tool using a common intermediate representation has to know the complete specification. It is even possible to strip away or add individual fields of commonly used types – independent of temporal, physical or lexical order.
- the coding is platform and language independent
- the coding offers a maximum of downward **and** upward compatibility
- a programmer is communicating through a generated interface, which allows programmers, knowing nothing about the SKILL, to interact with it, and ensures type safety. It also allows programmers to write tools in the language they know best⁷.
- stored data, that is never needed by a tool, will never be touched.
- new objects can be added to a file by appending data to the existing file.

Any of the arguments above have already been made in various contexts (e.g. [TDB⁺06] §13.13, [LA13], [xml06], [Lam87]), but there is, to the best of our knowledge, no solution bringing all these demands together into a single product that does the job automatically.

2 Outline

3 Related Work

To do (8)

To do (9)

There are many approaches similar to ours, but most of them have a different focus. This section shall provide a concise list of related approaches. For potential users of

⁵I.e. regular references and annotations.

⁶I.e. it can be distributed over many files.

⁷This is a problem especially in the scientific community, where many researchers work on similar problems but with completely different tools.

SKiLL, this might also present alternatives superior for individual use cases.

XML

XML is a file format (defined in [xml06]). The main differences are:

- + XML can be manipulated with a text editor⁸.
- + It is easier to write a libXML for a new language than to write a SKiLL back-end⁹.
- XML is not an efficient encoding in terms of (disk-)space usage. This can be overcome by the Efficient XML Interchange Format (EXI) (see [SK11]).
- XML is not type safe. This can be overcome partially by the XML Schema Definition Language (XSD).
- XML does not provide references to other objects out of the box.
- XML stores basically a tree, whereas a SKiLL file contains an arbitrary amount of graphs of objects.
- XML is usually accessed through a libXML, whereas SKiLL provides an API for each file format, thus a SKiLL user does not require any SKiLL skills, i.e. no knowledge about SKiLL types or the representation of serialized objects is required in general. To be honest, there are some language bindings, mainly for Java, which offer this benefit for XML as well.
- XML-files can not be appended with new data.

XML Schema definitions

SKiLLs description language itself is more or less equivalent to XML schema definition languages such as XSD (as described in [GSMT⁺08, PGM⁺08]). The most significant difference is caused by the fact that XML operates on trees and SKiLL operates on arbitrary graphs.

The type systems offered by SKiLL and XSD are quite different, thus it might be worth a look which one better fits ones needs.

JAXP and xmlbeansxx

For Java and C++, there are code generators that turn an XML schema file into code, which is able to deal with an XML in a similar way as proposed by this work. In case of Java the mechanism is even part of the standard library. The downside is that, to the best of our knowledge, this is only possible for Java and C++, thus it leaves us with portability issues. A minor problem of this approach is the lack of support for comment generation and the inefficient storage of serialized data.

XML based Approaches

There are various XML based formats. All XML based formats share the basic performance overhead of XML. <http://www.uni-koblenz-landau.de/koblenz/fb4/ist/rgebert/research/graph-technology>

⁸Whereas SKiLL files are binary and require a special editor, which will be provided by us eventually.

⁹This is only a relevant point if no bindings exist for the language you want to use.

To do (10)

To do (11)

To do (12)

To do (13)

ASN.1

Is not powerful enough to fit our purpose.

IDL

The published format is stated to be ASCII ([Lam87] §2.4), which will cause similar efficiency problems as raw XML does, if large amounts of data are stored.

To do (14)

Apache Thrift & Protocol Buffers

Thrift states that there is no sub-typing (see [Apa13]¹⁰). Protocol Buffers (see [Goo13]) seem not to support sub-typing either. Both seem to be a pragmatic approach to generalization of efficient network protocols. The type system of Protocol Buffers is also a rather pragmatic solution offering types such as unsigned 32 bit integers, which can not be represented in an efficient and safe way by e.g. Java. Both do not have storage pools, which are the foundation of our serialization approach and an absolute requirement for some of our optimizations, such as hints (see section 7.2).

Protocol Buffers provide a variable length integer type, namely `int64`, which seems to be binary compatible¹¹ with the variable length integer type used in SKill (see section C).

To do (15)

Solutions mentioned in this section have an advantage in terms of performance but offer little to no change tolerance at all. Therefore, they are best used in an environment with a lot of communication and total control over protocol changes. If protocol changes are frequent, the change tolerance provided by the SKill method may save a significant amount of development costs.

To do (16)

To do (17)

To do (18)

Java Bytecode, LLVM/IR and others

Although Java Bytecode (see [LYBB13]) and the LLVM Intermediate Representation (see [LA13]) are hand crafted formats, they served as a guiding example in many ways.

Language Specific Serialization

Language specific serialization is language specific and can therefore not be used to interface between subsystems written in different programming languages, without a lot of effort. Our aim is clearly a language independent and easy to use serialization format.

SKill Implementations of past Revisions

This revision is based on experience with previous versions of skill. Implementations can be found under . Descriptions of implementations can be found in .

To do (19)

To do (20)

To do (21)

To do (22)

¹⁰In section "structs", first sentence: "Thrift structs define a common object – they are essentially equivalent to classes in OOP languages, but without inheritance.", as of 29.Aug.2013

¹¹ The Protocol Buffer implementation seems not to optimize away the ninth flag, thus it might use an additional byte for very large numbers.

4 Syntax

This section discusses the syntax of the description language in brevity. The semantics is discussed in section 5, the file format is discussed in section 10.

We use the tokens `<id>`, `<string>`, `<int>`, `<float>` and `<comment>`. They equal C11-style¹² identifiers, strings, integer literals, float literals and comments respectively. Identifiers, strings and comments are explicitly enriched by printable Unicode characters above `\u007f`, although this feature should be used with care. For the sake of portability Unicode characters are restricted to 16 bit code points only. Usage of code points above `\uffff` has no defined behaviour¹³. We use a comment token, because we want to emit the comments in the generated code, in order to integrate nicely into the target language's documentation system.

4.1 Reserved Words

The language itself has only the reserved words **annotation**, **auto**, **const**, **include**, **with**, **bool**, **namespace**, **map**, **list** and **set**.

However, it is strongly advised against using any identifiers which form reserved words in a potential target language, such as Ada, C++, C#, Java, JavaScript or Python, as well as the identifiers **skillid**, **internal** and **api**.

4.2 The Grammar

A simplified overview of the grammar of a SKILL definition file is given below. For the sake of readability, fancy types, change modifiers and comments have been simplified or removed. The complete grammar can be found in appendix A. The detailed explanation of individual syntactic constructs will make use of the complete grammar.

To do (23)

To do (24)

```
UNIT :=
    TrueComment*
    INCLUDE*
    DECLARATION*

TrueComment := ("#" ~[\n]* \n)*

INCLUDE :=
    ("include"|"with") <string>+

DECLARATION :=
    DESCRIPTION
    <id>
    ((":"|"with"|"extends") <id>)?
    "{" FIELD* "}"

FIELD :=
    DESCRIPTION
```

¹²[ISO11] Annex D

¹³A code generator is expected to reject specifications containing these characters, whereas implementations are expected to treat user strings containing 32 bit code points correctly.

```

(CONSTANT|DATA) ";"

DESCRIPTION :=
  <comment>?
  (RESTRICTION|HINT)*

RESTRICTION :=
  "@ " <id> "(" (R_ARG (" " R_ARG)*)? ")"?

R_ARG := (<float>|<int>|<string>)

HINT := "!" <id>

CONSTANT :=
  "const" TYPE <id> "=" <int>

DATA :=
  "auto"? TYPE <id>

TYPE :=
  ("map" MAPTYPE
   | "set" SETTYPE
   | "list" LISTTYPE
   | ARRAYTYPE)

MAPTYPE :=
  "<" GROUNDTYPE (" " GROUNDTYPE)+ ">"

SETTYPE :=
  "<" GROUNDTYPE ">"

LISTTYPE :=
  "<" GROUNDTYPE ">"

ARRAYTYPE :=
  GROUNDTYPE
  ("[" (<int>)? "]" )?

GROUNDTYPE :=
  (<id>|"annotation")

```

Note: The Grammar is LL(1).¹⁴

To do (25)

To do (26)

4.3 Examples

Listing 1: Running Example

¹⁴In fact it can be expressed as a single regular expression.

```

/** A location in a file pointing to a character in that
    file. Assumes ordinary text files. */
Location {

    /** the line of the character starting from 0 */
    i16 line;

    /** the column of the character starting from 0 */
    i16 column;

    /** the file containing the location */
    File path;
}

/** A range of characters in a file. */
Range {

    /** first character; inclusive */
    Location begin;

    /** last character; exclusive */
    Location end;
}

/** A hierarchy of file/directory names. */
File {

    /** Name of this file/directory. */
    string name;

    /** NULL iff root directory. */
    File directory;
}

```

Includes, self references

Listing 2: Example 2a

with "example2b.skill"

```

A {
  A a;
  B b;
}

```

Listing 3: Example 2b

with "example2a.skill"

```
B {  
    A a;  
}
```

which is equivalent to the file:

Listing 4: Example 2

```
A {  
    A a;  
    B b;  
}  
  
B {  
    A a;  
}
```

Subtypes

Types can be extended using subtyping:

Listing 5: Subtyping Example

```
with "runningExample . skill"  
  
/** a message is just a string */  
Message {  
    string message;  
}  
  
/** located messages contain a location as well */  
LocatedMessage extends Message {  
    Location location;  
}
```

Containers

Container types can be used to store more elaborate data structures, then just plain values or references. In the current version, there is support for sets, maps, lists and arrays:

Listing 6: Containers

```
/** E.g. a user in a social network. */  
User {  
    string name;  
  
    /** friends of this user */  
    List<User> friends;  
}
```

```

    /** default values of permissions can be overridden on
        a per-user basis. The value is stored explicitly to
        ensure that the override survives changes of the
        permissions default value. */
    Map<User, Permission, Bool> permissionOverrides;
}

Permission {
    string name;
    bool default;
}

```

Unicode

The usage of non ASCII characters is legal, but discouraged.

Listing 7: Unicode Support

```

Ä {
    Ä v;
    Ä €;
}

```

4.4 Style Guide

This section provides some general hints on how to write readable specifications. A uniform appearance of specifications will improve the readability and therefore reduce development time.

Use speaking names Although type and field names are serialized, the overhead is a small constant. Furthermore, speaking names tend to prevent name clashes and decrease development time.

Use camelCase identifiers consisting of letters a-z only This rule enables code generators to create readable names for target languages. Letters a-z can be represented in most encodings, are used by the English language and can therefore assumed to be safe. Usage of other characters may require unreadable escape sequences.

If Acronyms have to be used followed by a word then the last capitalized character is interpreted as the first character of the subsequent word.

Use Capitalized user type names This convention is used by most programmers. If this convention contradicts the style guide or rules of a language, the language specific binding generator is expected to change casing according to the respective style guide.

Use lowercase field type names Field Types should consist of a single word only. This convention is used by most programmers. If this convention contradicts the style guide or rules of a language, the language specific binding generator is expected to change casing according to the respective style guide.

Indent Units with two spaces Specifications are likely to be processed with regular text editors. The lack of control flow allows for a small indentation level.

Comments, Restrictions and Hints should reside in their own lines Using separate lines for type or field modifications causes the amount of modification to be reflected in the visual appearance of the entity.

Separate field declarations with comments, restrictions or hints with one blank line from other field declarations This rule turned out to be very usable. If neither comments, hints or restrictions are used, the specification may choose to skip blank lines. Not using comments is discouraged. Separating subsequent type definitions with two blank lines turned out to be useful as well.

Declare super types first The skill specification language itself enforces no order on type declarations. Human readers however may not be able to read specifications where super types are used before they are specified. This rule does not apply to types used as part of field declarations.

Split the specification into logical units Specifications distributed into several files consisting of small logical units are more accessible using just a text editor than a single large file. Splitting up a specification enables tool builders to omit parts irrelevant for their specific tool.

5 Semantics

This section will describe the meaning of specifications by explaining the effect of declarations.

5.1 A Specification File

```
UNIT :=  
  INCLUDE*  
  DECLARATION*
```

Specification files are fed into binding generators, which generate code that provides means to deal with instances of the declared types.

SKILL specifications consist of a set of declarations which themselves consist of fields. A declaration is roughly equivalent to a type declaration in an object oriented programming language. The main difference is, that declarations are pure data, because we do not offer a real execution model. The only operations from the perspective of SKILL are loading and storing of data.

A declaration will instruct the language binding generator to create a type which has the declarations name and consists of the fields specified in the body of the declaration. Fields behave just like fields in object oriented programming. Both form the structure of serialized data and are identified using human readable names.

Types are discussed in section 6. Restrictions and Hints are discussed in section 7.1 and 7.2.

To do (27)

5.2 Includes

```
INCLUDE :=  
  ("include"|"with") <string>+
```

Includes are used to structure a specification into smaller models, e.g. by moving data that is only used by some tools to its own file.

The files referenced by the `include` statement are processed as well. The declarations of all files transitively reachable over `with` statements are collected, before any declaration in any file is evaluated. The order of inclusion is irrelevant. The same file may even be included multiple times by the same `include` statement.

Therefore evaluation of declarations happens as if all declarations were defined in a single file.

5.3 Type Declarations

```
DECLARATION :=  
  DESCRIPTION  
  <id>  
  ((":"|"with"|"extends") <id>)?  
  "{" FIELD* "}"
```

The SKILL specification language is all about type declarations. A type declaration consists at least of a name and a body, containing field declarations.

Descriptions

```
DESCRIPTION :=  
  COMMENT?  
  (RESTRICTION|HINT)*
```

Type (and field) declarations can be enriched with descriptions.

Comments provided in the SKILL specification will be emitted into the generated code¹⁵ to serve as a natural language description of the respective entity. This approach enables users to get tool-tips in an IDE showing him this documentation. Future revisions are expected to expand the notion of a comment by introducing tags as used by most documentation generation systems, such as doxygen [vH13] or javadoc [jav13].

Comments Comments will ignore preceding whitespace and `/**` tokens. Comments end with `*/`. Comments can not be nested. Comments start with a text and can be followed by tags. Valid tags are

- `@see type/field`
- `@deprecated message`
- `@author name`
- `@version string`

¹⁵ If the target language does not allow for C-Style comments, the comments will be transformed in an appropriate way.

- *@note message*
- *@todo message*

These tags are parsed separately and provided to the back end as a list, to allow the back end to convert them to the format required by the dominating documenting system. The behaviour in the presence of other tags is unspecified. Ill-formed tags must not cause a binding generator to fail.

If a valid type or field is used as argument to see, the type is translated by the back end to match the generated type or field.

Line breaks and empty lines are dealt with as known by \LaTeX : Empty lines separate paragraphs, while ordinary line breaks will be ignored.

Restrictions and hints will be explained below.

To do (28)

To do (29)

To do (30)

To do (31)

To do (32)

Subtypes

A subtype of a user type can be declared by appending the keyword `with` (or `:` or `extends`) and the supertypes name to a declaration. A subtype behaves like a subtype in object oriented programming. Subtypes inherit all fields of their super types. Only user defined types can be sub-typed. In order to be well-formed, the subtype relation must remain acyclic and must not contain unknown types.

5.4 Field Declarations

```
FIELD :=
  DESCRIPTION
  (CONSTANT|DATA) ";"
```

```
CONSTANT :=
  "const" TYPE <id> "=" <int>
```

```
DATA :=
  "auto"? TYPE <id>
```

Types are sets of named fields. Fields are either constants or real data.

A usual field declaration consists of a type and a field name. In this case, the field declaration behaves like a field declaration in any object oriented programming language, except that the field data will be serialized.

5.4.1 Constants

A `const` field can be used in order to create guards or version numbers. The deserialization mechanism has to report an error if a constant field has an unexpected value. This mechanism is intended to be used basically for preventing from reading arbitrary files and interpreting them as the expected input. The mechanism can be used defensively, because storing constant fields creates a constant overhead and is not influenced by the number of instances of a type.

Only integer types can be used as constants.

5.4.2 Transient Fields

Transient fields, i.e. fields which are used for computation only, can be declared by adding the keyword `auto` in front of the type name. The language binding will create a field with the given type, but the content is transparent to the serialization mechanism. This mechanism can be used to add fields to a data structure, which simplify algorithms computing the interesting data, if these helper fields are not of interest after computation. This is especially useful in combination with the possibility to add or drop some fields while generating the binding for a specific tool. The mechanism can also be used for a field with content that can likely be computed very fast.

The keyword `auto` is used, because the content of the field is computed automatically. Besides the name, it has nothing to do with the `auto` type declaration of C++.

To do (33)

5.4.3 Change-Modifier

Change modifiers can be applied to fields or whole type declarations. Their intended usage is to enable users to specify the contents of their files along their tool-chain by adding (++) data to the specification or by removing (--) it. This is an important behaviour in a large tool chain, because it enables tools to have a lean API to the data required by them, without having to maintain a large amount of cloned file specifications.

Usage of ++ and -- on a type declaration is a shorthand for modifying all contained fields individually. Usage of == is only legal on whole type declarations and is considered to be the "ultimate truth", thus there can be at most one occurrence for each type.

Usage of change modifiers has an impact on the semantics of a specification. Assuming a type *T* has specifications and modifications, then

- There is exactly one type declaration for *T* without a change modifier.
- An arbitrary amount of ++ and - modified declarations with varying restrictions or fields can be applied to a type. A field can either occur with an arbitrary amount of ++ or -- modifications, but it must not have both.
- There can be at most one == modified type declaration. In that case all other declarations are ignored.

To do (34)

5.5 Field Types

```
TYPE :=
  ("map" MAPTYPE
   | "set" SETTYPE
   | "list" LISTTYPE
   | ARRAYTYPE)

MAPTYPE :=
  "<" GROUNDTYPE ("," GROUNDTYPE)+ ">"

SETTYPE :=
  "<" GROUNDTYPE ">"
```

```
LISTTYPE :=
  "<" GROUNDTYPE ">"
```

```
ARRAYTYPE :=
  GROUNDTYPE
  ("[" (<int>)? "]" )?
```

```
GROUNDTYPE :=
  (<id>|"annotation")
```

Basic types are just an identifier with the type name. For compatibility reasons¹⁶, type names are case insensitive.

Types are explained in-depth in section 6.

5.5.1 Container Types

The type system has a built-in notion of arrays, maps, lists and sets. Note that all of them are, from the view of serialization, equivalent to length encoded arrays. Their main purpose is to increase the usability of the generated Application Programming Interface (API).

These containers showed to increase the usability and understandability of the resulting code and file format.

5.5.2 Annotations

The `annotation` type is basically a typed pointer to an arbitrary user type. Its main purpose is to provide extension points in the form of references to objects, who's type could not be known at the time of the specification of the annotation field.

5.6 Type Annotations

```
RESTRICTION := "@" <id> "(" (" (R_ARG (" , " R_ARG)*)? ")" )?
```

```
R_ARG := (<float>|<int>|<string>)
```

```
HINT := "!" <id>
```

SKILL offers two kinds of type annotations: restrictions and hints. Restrictions can be used to restrict a type, e.g. by reducing the range of possible values of an integer fields to those above 23. Hints can be used to optimize the generated language binding.

Both are explained in-depth in section 7.1 and 7.2.

5.7 Name Resolution

Because SKILL is designed to be downward and upward compatible and offers subtyping, it is possible that a future revision of a file format specification will add a field with a name that already exists in a subtype. In general, it is assumed that the current

¹⁶ Some programming languages, e.g. Ada, do distinguish types by casing of identifiers. Using types which differ only in case in such languages would be very nasty, because type name would have to be escaped in some way. Furthermore, clashes with reserved words can usually solved by capitalization.

maintainer of the super class does not know about all subclasses. Thus, it is desirable to have a mechanism which ensures client code to work correct after such a change.

Therefore, identical field names are legal in SKILL and refer to different fields, as long as they belong to different type declarations.

A generated language binding has to provide means of accessing a field shadowed by a field of the same name in a subtype. In most languages there are built-in mechanisms for this task. Language binding generators shall take care that they do not override field access methods in a way that will actually make the field of the super type inaccessible.

6 The Type System

The description language and the file format are both intended to be *type safe*. The notion of type safety is usually connected to a state transition system. In our context, the only observable state transitions are from the on-disk representation to the in-memory representation and vice versa. Thus, with *type safe* we want to state that deserialization and serialization of data will not change the type of the data. It is further guaranteed that deserialized references will point to objects of the static type of the reference. Further, if one were to deserialize an object of an incompatible type, an error will be raised.

These properties require some form of platform independent type system¹⁷, which is described briefly in this section. The general layout of the type system is visualized in Fig. 1.

Common Abbreviations

We will use some common abbreviations for sets of types in the rest of the manual:

Let ...

... \mathcal{T} be the set of all types.

... \mathcal{U} be the set of all user types.

... \mathcal{I} be the set of all integer types, i.e. $\{\text{i8, i16, i32, i64, v64}\}$.

... \mathcal{B} be the set of all built-in types.

6.1 Built-In Types

The type system provides built-in types, which are the building blocks of type declarations.

Integers

Integers come in two flavors, fixed length and variable length. There is currently only a single variable length integer type, namely Variable length 64-bit signed integer (v64). The variable length integer type can store small values in a single byte (see appendix C for details). Large values ($\geq 2^{55}$) and negative values require one additional byte, i.e. nine bytes.

To do (35)

¹⁷In contrast to e.g. C, objects of certain type have a known length and endianness.

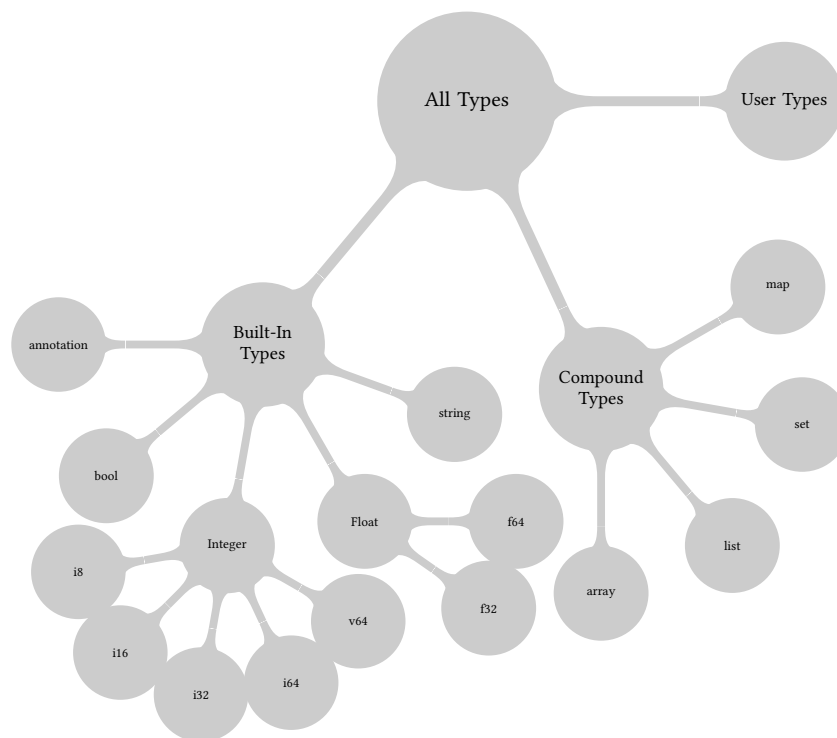


Figure 1: Layout of the Type System

Booleans

Booleans can store the values $\text{true}(\top)$ and $\text{false}(\perp)$. Unlike most C programmers, we do not perceive booleans as integers.

Annotations

Annotations are designed to be the main extension points in a file format. Annotations are basically typed pointers to arbitrary types. This is achieved by adding the type of the pointer to a regular reference. A language binding is expected to provide something like an annotation proxy, which is used to represent annotation objects. If an application tries to get the object behind the proxy for an object of an unknown type, this will usually result in an error or exception¹⁸. Therefore language bindings shall provide means of inspecting whether or not the type of the object behind an annotation is known.

To do (36)

To do (37)

To do (38)

Strings

Strings are conceptually a variable length sequence of utf8-encoded unicode characters. The in-memory representation will try to make use of language features such as `java.lang.String` or `std::u16string`. The serialization is described in section 10. If

To do (39)

To do (40)

¹⁸The reflection mechanism allows for other solutions, but raising an exception is the most obvious reaction.

a language demands NULL-termination in strings, the language binding will ensure this property.

Strings should not contain NULL characters, because this may cause problems with languages such as C. However, SKill is NULL character agnostic, thus no guarantees are made.

The API shall behave as if strings were defined with the hint `pure` and `unique` (see section 7.2).

To do (41)

NULL Pointer

Fields of type strings, annotations or a user type can store a NULL pointer. These types are by default nullable as NULL pointers are their default values. Fields of these types can be declared nonnull using the `nonnull` restriction (see Section 7.1.2.1).

Floating Point Numbers

For convenience, it is possible to store 32 bit and 64 bit IEEE-754 floating point numbers. For a description, see [jee08] especially §3.4.

6.2 Compound Types

The language offers several compound types. Sets, Lists and variable length Arrays¹⁹ are basically views onto the same kind of serialized data, i.e. they are a length encoded sequence of instances of the supplied base type. Arrays are expected to have a constant size, i.e. they are not guaranteed to be resizable. Sets are not allowed to contain the same object twice. All compound types will be mapped to their closest representation in the target language, while preserving these properties. Maps are viewed as a representation of serializable partial functions. Therefore they have two or more type arguments.

To do (42)

To do (43)

6.3 User Types

To do (44)

User types can be interpreted as sets of type-name-pairs. Built-in types can be wrapped in order to give them special semantics. For example an appointment can be represented as:

To do (45)

To do (46)

Listing 8: Example User Type

```
Appointment {
  /** seconds since 1.1.1970 0:00 UTC. */
  i64 time;

  /** A topic, such as "team meeting". */
  string topic;

  /** the name of the room. */
  string room;
}
```

To do (47)

¹⁹I.e. arrays, which do not have constant size. Constant length arrays exist as well.

To do (48)

6.3.1 Legal Types

The given grammar of SKill already ensures that intuitive usage of the language will result in legal type declarations. The remaining aspects of illegal type declarations boil down to ill-formed usage of type and field names and can be summarized as:

- Field names inside a type declaration must be unique. Field names of super types are not relevant.
- The subtype relation is a partial order²⁰ and does not contain unknown types.
- Any base type has to be known, i.e. it is a user type defined in any document transitively reachable over include commands.
- The type names must be unique in the context of all²¹ types.

To do (49)

6.3.2 Equivalence of Type and Field Names

Type and field names, i.e. any strings referenced from reflection data stored in a SKill file, shall be treated as equivalent, if they were equal after converting all characters to lower case.

We recommend using CamelCase in SKill definition files in order to provide a hint to the language binding generator on how to separate parts of identifier names. For example an Ada generator will add underscores to the names in the generated interface leading to a more natural feeling for Ada programmers. In order to influence this behaviour, any SKill specification front-end shall provide an option to interpret a single '_' as explicit word separation and double underscores as escape for a single underscore. In this mode, a single underscore will cause a Java generator to use CamelCase.

6.4 Fancy Types

This section will explain fancy types that are made visible to the user, although they do not exist in the binary file per se.

All fancy types can be mapped to equivalent user type definitions.

TBD

6.4.1 Interfaces

TBD

To do (50)

To do (51)

6.4.2 Enums

To do (52)

TBD

To do (53)

6.5 Views

To do (54)

```
'view' (<id> '::')* <id> '.' <id> 'as'  
<field-declaration>
```

²⁰In fact it forms a forest.

²¹From the perspective of a client, i.e. all types that were declared at the time of the generation of its interfaces plus all types that are ever observed in the form of *unknown* types encoded in SKill files.

Das hier sollte genutzt werden um die shadowing convention zu ändern. Tatsächlich ist die Namensauflösung dann ziemlich kompliziert.

Views gehören definitiv in den full bereich.

Der Typ der field declaration muss ein subtyp der angesehenen declaration sein.

Es dürfen weitere restrictions und hints vergeben werden. Es gelten bei restrictions beide und bei hints im zweifelsfall die des views.

Falls ein subtyp ein feld des supertyps mit selbem namen ohne view überschreibt, dann muss es einen view geben, der das feld zugänglich macht.

Views dürfen den selben namen haben, wie das gesehne feld.

Die serialisierung eines views ist konstant teuer und hat die ID 31.

Man muss hier sehr deutlich sagen, dass dieser mechanismus der beseitigung historischer fehlentscheidungen dient und mit äußerster vorsicht zu benutzen ist.

Views sollten einen !export hint enthalten, der dafür sorgt, dass der typ in der exportierten datei entsprechend des views konvertiert wird. Dadurch sind historische fehlentscheidungen bei der namensgebung nicht mehr ganz so in stein gemeißelt.

To do (55)

6.6 Typedefs

To do (56)

COMMENT

```
'typedef' <id> (<restriction>|<hint>)* <id> ';' ;'
```

Creates a type name as the restricted and hinted version of another name. This feature is handy in large specifications. Typedefs can be provided with a comment, that is made accessible to the back-end.

The front-end shall provide a mechanism to project away all typedefs.

Typedefs can syntactially be used to nest containers inside of containers. This is forbidden for coding reasons and will therefore result in an error. For usability reasons using containers in typedefs and typedefs in container definitions is ruled to be legal.

In case of nested containers resulting from typedefs, an error message shall be created at generation time stating that non ground type container argumet resulting from typedef blabla.

Typedefs are syntactic sugar which is not exported to the binary file and which shall be projected away to code generators on demand²².

6.7 Default Values

Default values are:

Type	Value
integers	0
floats	0.0
enums	first definition
containers	∅
others	null

6.8 Examples

This section will present some examples of ill-formed type declarations and brief explanations.

²²This is required in order to keep basic implementations at a sane cost.

Listing 9: Legal Super Types

```
EncodedString extends string {  
    string encoding;  
}
```

Error: The built-in type “string” can not be subclassed.

Listing 10: Legal Type Names

```
/*  
    The german word for "car".  
*/  
Auto {  
    ...  
}
```

Error: “Auto” is a reserved word and can not be used as type name.

Listing 11: Usage of Unknown Types

```
A {  
    map<A, B> f  
}
```

Error: The field “A.f” refers to a missing type B. Did you forget to include “B.skill”?

To do (57)

To do (58)

To do (59)

To do (60)

7 Type Annotations

SKILL provides two concepts of extending the basic type system used in the serialization process.

The first concept is called restriction and is inspired by the concept of (type or class) invariants. This concept can be used to restrict the set of legal objects storable in a field or the set of legal instances of a class.

The second concept is called hint. Hints are used to improve the generated language binding and do not influence types per se.

To do (61)

7.1 Restrictions

Restrictions can be added to type declarations and fields. They can occur in any number at the same places as comments. Restrictions start with an @ followed by a name and optional arguments. If multiple restrictions are used, the conjunction of them forms the invariant, i.e. all of them have to apply. Field and type restrictions do not share a common ID-range and are explained their respective subsections.

If Restrictions are used on compound types, they expand to the components of the respective compound type. Restrictions can not be combined with map-typed fields.

Restrictions are checked by the generated binding at least before serialization and after deserialization. If checking restrictions involves fields, which are not present in a deserialized file, the respective restriction is ruled to hold. This is important to guarantee compatibility with older or newer versions of a file format used in a tool chain. This behavior puts the burden of fulfilling restrictions to the creator of data.

Restrictions are serialized to ensure the asserted properties. The serialization mechanism and an optional recovery strategy is specified alongside each restriction definition. Restrictions with even IDs may be ignored by a binding not implementing the restrictions. These restrictions must implement the specified recovery strategy. Restrictions with even IDs are serialized by the function $\llbracket ID \rrbracket_{v64}$ ²³.

Reading a field that does not comply to the specified or serialized restrictions renders the field partial (see section 118).

To do (62)

To do (63)

7.1.1 Type Restrictions

To do (64)

Type restrictions can be applied to either **any** type definitions or **base** type definitions. They extend to their sub-types automatically.

To do (65)

To do (66)

To do (67)

7.1.1.1 Unique

Objects stored in a storage pool have to be distinct in their serialized form, i.e. for each pair of objects, there has to be at least one field with a different value. Because the combination of unique with sub-typing has counter-intuitive properties, we decided that using the unique restriction together with a type that has sub- or super-types is considered an error, which has to be detected at runtime.

Unique implies an immutable in-memory representation. If the representation were allowed to be mutable, the objects might require a merge after arbitrary operations, which comes at huge runtime costs. Therefore implementations shall provide a builder for unique instances, as well as a means of deleting existing instances upon write operations.

Furthermore, be warned that adding or removing fields from unique types can cause serious compatibility issues.

ID	0
Applies to	any
Serialization	ε
Recovery	unify duplicates and add restriction to output

Listing 12: Examples

```
@unique Operator {
    string name;
}
@unique Term {
    Operator operator;
    Term[] arguments;
}
```

7.1.1.2 Singleton

There is at most one instance of the declaration. Singletons must not have sub-types. On the other hand, singletons may have super types. In fact instances of enums are represented by singleton subtypes of the enum types. This enables enum instances to have fields.

²³The binary file would not be parsable otherwise.

ID	1
Applies to	any
Serialization	ε
Recovery	report an error, if more then one instance exists, otherwise add to output

Listing 13: Examples

```
/* * Stores properties of the target system. */
@singleton System {
    string name;
    ...
}
```

Listing 14: Enum like

```
@abstract Weekday {
    string name;
    i8 index;
}
@singleton Monday : Weekday {
}
@singleton Tuesday : Weekday {
}
...
```

is, besides the generated API, equivalent to

Listing 15: Enum like

```
enum Weekday {
    Monday, Tuesday, ...;

    string name;
    i8 index;
}
```

7.1.1.3 Monotone

Instances of this type can not be deleted. Furthermore instances can not be modified once they have received a SKILL ID, because they might have been serialized already. This restriction is basically a type system way of ensuring properties required by fast append operations. It should be used, where ever it is not necessary to delete instances of a type. The monotone restriction can only be added to base types and expands to all sub types of the base type²⁴.

Monotone types can be treated by a language binding in a very optimized way.

²⁴ This behavior is caused by the fact, that for $A \leq B$, all B instances are As as well. If B would not be monotone, deleting a B would directly delete an A. If A would not be monotone, deleting an A might delete a B.

The monotone restrictions implies the monotone hint (see section 7.2.11).

ID	2
Applies to	base
Serialization	ε
Recovery	add to file

Listing 16: Examples

```

/* * just kidding */
@Monotone SocialNetworkPost {
    string message;
}

/* * this is monotone as well */
PrivatePost extends SocialNetworkPost {
    string recipient;
}

```

7.1.1.4 Abstract

The restricted type must not have static instances. This is similar to the abstract classes in C++ or Java. In contrast to C++/Java abstract, super types can be non-abstract.

Note This restriction does not apply to sub-types.

ID	3
Applies to	base
Serialization	ε
Recovery	raise error, if static instances exist, add to file otherwise

7.1.1.5 Default

7.1.2 Field Restrictions

To do (68)

Filed restrictions use the type α to denote the fields type.

To do (69)

7.1.2.1 NonNull

Declares that the argument field cannot be NULL.

Note that fields, which have not been initialized, contain NULL values if they have no other default specified explicitly.

Cannot be used on annotation.

ID	0
Applies to	String, Usertypes
Signatures	nullable:
Serialization	ε
Recovery	add to file

Listing 17: Examples

```

Node {
  /* null-edges are pointless */
  @nonnull Node[] edges;
}

```

7.1.2.2 Default

integer und floats sind ihre entsprechenden werte, solange sie in Range liegen. strings
können string literale sein pointer artige können namen von Singletons eines passenden
typs sein.

ID	1	To do (70)
Applies to	ground types	
Signatures	<code>default(value): α</code>	To do (71)
Serialization	<code>[[value]]_{α}</code>	To do (72)
Recovery	add to file; replace serialized default, if one exists; raise error if default value is invalid ²⁵ .	To do (73)

Listing 18: Examples

```

/* * a multi-graph with a weighted edges and labeled
   nodes. */
Graph {
  Node[] nodes;
  Edges[] edges;
}

Node {
  @default("")
  string label;
}

Edge {
  Node from;
  Node to;

  @default(1.0)
  f32 weight;
}

```

7.1.2.3 Range

Range restrictions are used to restrict ranges of integers and floats. They can restrict the minimum or maximum value or both. Restrictions can be inclusive or exclusive – the default is inclusive.

Note that this will change the implicit default value of the argument field to *min* iff $0 \notin [min, max]$. If an explicit default value has been specified, it must not contradict the range.

ID	3
Applies to	Integer, Float
Signatures	$\text{range}(\text{min}, \text{max}, \text{boundaries}): \alpha \times \alpha \times \text{string}^?$ $\text{min}(\text{min}, \text{boundaries}): \alpha \times \text{string}^?$ $\text{max}(\text{max}, \text{boundaries}): \alpha \times \text{string}^?$
Serialization	$\llbracket \text{min}_{\text{inclusive}} \rrbracket_{\alpha} \circ \llbracket \text{max}_{\text{inclusive}} \rrbracket_{\alpha}$
Recovery	add to file; intersect with serialized range; raise an error, if intersection is empty

Listing 19: Examples

```

RangeRestricted {
  @min(0)
  v64 natural;

  @min(1)
  v64 positive;
  /* or */
  @min(0, "exclusive")
  v64 positiveAlt;

  @range(0.0, 360.0, "inclusive, _exclusive")
  f32 angle;
}

```

7.1.2.4 Coding

The arguments field data chunks are encoded using the argument coding. No codings are specified by SKILL. Potential arguments include "zip" or "lzma" or some sort of encrypting container. To do (??)

Usage of codings is discouraged, because of the burden it puts on implementing a binding generator for some combinations of programming language and platform and the implied loss of portability.

Usage of coding might be necessary to fulfill legal obligations in a specific use case. To do (75)

Note Coding is designed in a way, that allows skipping fields if the coding is unsupported. If a coding requires additional information, this information has to be placed somewhere in the field data or in a side-channel of the application, such as another file or a client-server communication.

ID	5
Applies to	any
Signatures	$\text{coding}(\text{name}): \text{string}$
Serialization	$\llbracket \text{name} \rrbracket_{\text{string}}$
Recovery	add to file on write; if unsupported, treat field ignore-hinted and the type as monotone-restricted

Listing 20: Coding Examples

```

ToolDescription {
    /** the log is written by a tool, but usually unused
        afterwards */
    !lazy
    @coding("zip")
    string log;
}

```

7.1.2.5 Constant Length Pointer

To do (76)

The argument pointer is serialized using i64 instead of v64. This can be used on regular references and annotations. This can be combined with the coding restriction. The restriction makes only sense if the generated binding supports lazy reading of partial storage pools and if the files that have to be dealt with, would not fit into the main memory of the target machine. Using this restriction will most certainly increase the file size and does not restrict any pointer targets.

This restriction is serializable and, thus, does not affect compatibility in any way.

ID	7
Applies to	String, Annotation, Usertypes
Signatures	constantLengthPointer
Serialization	ε
Recovery	add to file on write; if unsupported, treat field ignore-hinted and the type as monotone-restricted; on append, use whatever the file specified.

Listing 21: Examples

```

/* stored points to information may exceed the available
   main memory, thus we have to access it directly from
   disk */
PointsToTargets {
    @constantLengthPointer
    Context context;
    @constantLengthPointer
    HeapObject object;
    @constantLengthPointer
    PointsToSet targets;
}

```

7.1.2.6 OneOf

OneOf restrictions are used to restrict potential targets of annotations. Restrictions are still nullable. This is the foundation for tagged union types.

ID	9
Applies to	Integer, Float
Signatures	<code>oneOf (type1, ..., typeN): $\mathcal{U} \times \dots \times \mathcal{U}$</code>
Serialization	<code>[[type₁, ..., type_n]]_{TYPE}</code>
Recovery	add to file; intersect with serialized range; raise an error, if intersection is empty

7.2 Hints

Hints are annotations that start with a single `!` and are followed by a hint name. Hints are used to optimize the behavior of the generated language binding. They do not impact the semantics of type declarations or stored data. Therefore they will not be serialized.

Language bindings shall provide the same public interface as if no hints were used.

Language binding generators shall provide an option that adds a hint to all applicable declarations.

7.2.1 Owner

Can be used on: Base type declarations.

Arguments: Names of tools that own the type hierarchy, i.e. that can modify instances of the annotated base type.

Note: At generation time, the binding generator needs to know tool names, that are owned by the generated binding, in order to hard-link this behavior.

Note: Arguments are lists of names, in order to allow to separate the forest of types into different areas of concern.

Types without this hint are owned by every binding. If a binding is created without ownership information, it owns all types implicitly. These two rules are required in order to keep ownership optional, even if annotations exist.

Note on implementation: All types that are not owned by the tool are marked as `ReadOnly`.

Listing 22: Owner Example

```
!owner (ColorOwner)
Color { i8 r; i8 g; i8 b; }

/*
 * A GraphProvider should be a ColorProvider as well, but
 *   if it's not, it can still set existing
 *   colors. A ColorProvider on the other hand is able to
 *   create new colored graphs.
 */
!owner (ColorOwner, GraphOwner)
Node {
    Color color;
    set<Node> edges;
}
```

7.2.2 Provider

Can be used on: Type or field declarations.

Arguments: Names of the tools that provide the target field.

Note: If used on type declarations, it extends to all field declaration inside of it and all sub-types.

Note: If used on a field declaration, which belongs to a type declaration, that is not provided by the same provider, the provider of the surrounding type does not provide the field.

Note: A Provider name has to be specified at tool-generation time.

The target field can not be rendered partial by a write or append operation.

Listing 23: Provider Example

```
!provider( ColorProvider )
Color { i8 r; i8 g; i8 b; }

/**
A GraphProvider creates graphs.
A ColorProvider adds color.
*/
!provider( GraphProvider )
Node {
    !provider( ColorProvider )
    Color color;
    set<Node> edges;
}
```

7.2.3 Remove Restrictions

Can be used on: Anywhere

Arguments: Optional restriction names; if none is supplied, all restrictions are removed; if "unknown" is specified, only restrictions unknown to the generated binding are removed

Removes restrictions obtained from a file instead of keeping and adhering to them. If a restriction kind is to be removed but the same kind is specified, the specified kind will be present in the output. This behaviour is required in order to replace some restrictions with arguments, such as range, because their recovery strategy may have undesired consequences.

Rationale: This Restriction is used to deal with decisions made in the past. This behavior is usually undesired, therefore it has to be enabled explicitly.

Listing 24: Remove Unknown Restriction Example

```
/** in the past, there was only one display, i.e. we
    want to remove the @singleton */
!removeUnknownRestriction
Display { }
```



```

/** From now on, we allow to specify a dedicated primary
    display */
@Singleton
PrimaryDisplay{
    Display current;
}

```

7.2.4 Constant Mutator

The restriction can be used to change a constant. This might be required in order to upgrade the version of a data set.

The file reader will accept values between *min* and *max* and write a new file with value *new*.

Note Appending is not possible if a constant changed. A binding shall silently fall-back to write.

Listing 25: Constant Mutator Example

```

/** From now on, we allow to specify a dedicated primary
    display */
@Singleton
ToolInfo{
    /** the guard stays the same */
    const i16 guard = 0xABCD;

    /** we accept files adhering to version 1 as well,
        because the format is compatible in one direction */
    !constantMutator(1, 1)
    const i16 version = 2;

    string toolName;
}

```

7.2.5 Mixin

The actual Implementation of UserTypes is left to the client. This is a very generator dependent feature.

To do (77)

7.2.6 Flat

Can be used on a type definition with a single field to indicate that the type definition shall be hidden in the exported API. This is necessary, if containers ought to be shared between multiple entities.

7.2.7 Unique

Can be used on: Type declarations.

Serialization shall unify objects with exactly the same serialized form. In combination with the @unique restriction, no error shall be reported on serialization.

To do (78)

Note that this will increase the runtime complexity of the serialization phase from $O(n)$ to $O(u \cdot n \log(n))$, where n is the size of the state and u is the number of types declared to be unique.

7.2.8 Pure

Can be used on: Type declarations.

Deserialized objects of the annotated type shall not be modifiable. The generated interface will provide a copy operation, which will create a modifiable copy of the object. An example of this behavior is the string pool. An equivalent, in terms of API observable behavior, would look as follows:

Listing 26: User Strings

```
! pure
! unique
UserString {
    ! OnDemand
    i8 [] utf8Chars;
}
```

7.2.9 Distributed

Can be used on: Field declarations.

A static map will be used instead of fields to represent fields of definitions in memory. This is usually an optimization if a definition has a lot of fields, but most use cases require only a small subset of them. Because hints do not modify the binary compatibility, some clients are likely to define the fields to be distributed or even lazy.

To do (79)

Note that this will increase both the memory footprint²⁶ and the access time for the given field and will only be a benefit for memory-cache locality reasons, because single objects can be significantly smaller²⁷. The internal representation will change from `o.f`, i.e. a regular field, to `pool.f[o]`, i.e. a map in the storage pool which holds the field data for each instance.

Note that the presence of distributed, lazy or ignored fields will require objects to carry a pointer to their storage pool, which may eliminate the cache savings completely²⁸.

7.2.10 OnDemand

Can be used on: Field declarations.

Deserialize the respective field only if it is actually used. `OnDemand` implies distributed. This hint should be used, if fields are usually not accessed, e.g. in the context of error reporting.

²⁶ Because additional data structures, such as trees, are required in order to provide acceptable access times.

²⁷ Note this does not apply to operations on distributed fields, but to operations on objects having distributed fields.

²⁸ As a matter of fact, this is not true in case of non object oriented languages such as C.

7.2.11 Monotone

Can be used on: Base Type declarations.

Instances of the argument type will not be deleted. New instances can be added to the state. This allows an optimized treatment of data, because assumptions about object IDs can be made.

Accessing lazy fields of instances of monotone types is about as efficient as accessing ordinary fields. Note that usage of v64 and references requires deserialization of all fields in a type block. Thus, it might be worth considering usage of constant length integers or the `constantLengthPointer` restriction, if this case is encountered on a regular basis.

In contrast to the monotone restriction, the hinted monotone property applies only to the generated binding.

7.2.12 ReadOnly

Can be used on: Base Type declarations.

The generated code is unable to modify instances of the respective type. This hint can be used to provide a consistent API while preventing from logical errors, such as modifying data from a previous stage of computation. The `ReadOnly` hint expands to all subtypes of the base type, because this is the only safe way²⁹.

`ReadOnly` implies `Monotone`. If a `ReadOnly` field is partial, an error has to be thrown while reading the file.

7.2.13 Ignore

Can be used on: Type and Field declarations.

The generated code is unable to access the respective field or any field of the type of the target declaration. A language binding shall raise an error (or exception), if the field is accessed nonetheless. This hint can be used to provide a consistent API for a combined file format, but restrict usage of certain fields, which should be transparent to the current stage of computation. This is actually more restrictive than deleting fields from declarations, because the generated reflective API will respect this hint.

`Ignore` overrides any other hint and a binding may ignore any restriction applying to an ignored field.

7.2.14 Pragma

Can be used on: anything.

Usage: `!pragma <ID>+`

Example: `!pragma tree`

The `pragma` is used to pass additional information to the code generator. This mechanism is intended to be a short term solution to ease development of the language.

²⁹This behavior is caused by the fact, that for $A <: B$, all B instances are A s as well. If A would not be `ReadOnly`, modifying an A would directly modify a B . If B would not be `ReadOnly`, modifying a B might modifying an A .

A common mistake that people make when trying to design something completely foolproof is to underestimate the ingenuity of complete fools.

Douglas Adams in Mostly Harmless

Part II

User Interface

8 In Memory Representation

To do (80)
To do (81)
To do (82)

This section is about the generated API provided to the programmer and the representation of objects inside memory. It is meant as a concise guideline for implementers of language binding generators.

The behaviour in case of a mismatch between the expected type of a field and the type stored in a file is left unspecified. This will be changed in a future version. For now only matching types can be assumed to be compatible and portable.

To do (83)
To do (84)
To do (85)
To do (86)

8.1 API

The generated API has to be designed in a way that integrates nicely with the language's programming paradigms. For example in Java it would be most useful to create a state object, which holds state of a bunch of serializable data and provides iterators over existing objects, as well as factory methods and methods to remove objects from the state object. The serialized types can be represented by interfaces providing getters and setters, using hidden implementations only visible for the state object.

To do (87)
To do (88)
To do (89)
To do (90)
To do (91)
To do (92)
To do (93)
To do (94)
To do (95)
To do (96)

8.2 Representation of Objects

The combination of laziness and consistency has the effect that representation of objects inside memory is rather difficult. This section describes data structures and algorithms which can do the job in a sufficiently efficient way. In this section, we assume that all fields are present as arrays of bytes. We will describe the effects of parsing fields in an unmodified state, in a modified state, how to modify a state and finally how to write a state back to disk.

8.2.1 Proposed Data Structures

A state has to contain at least

- an array of strings
- type information obtained from a file
- storage pools

To do (97)

A storage pool has to hold the images of (or references to) fields, which are not yet parsed.

Objects are required to have an ID field, which corresponds to the ID of the deserialized(!) state. This field is required in order to map the lazy fields to the correct objects. It can also be reused in the serialization phase to assign unique IDs, which will be used instead of pointers.

Objects of types with eager fields should have the respective fields. For example, the declaration `T {t a; !lazy t b;}` should be represented by an object

```

InternalTObject {
    long ID;
    t a;
    /* getA, setA... */

    StoragePool tPool;
    /* getB, setB... */
}

```

Note that the pointer to the enclosing storage pool is required for the correct treatment of lazy and distributed fields. This is the case, because the pool holds the field data.

Now that we have a representation of objects, we still have to organize storage of objects across storage pools. The possibility of inheritance requires a view onto stored objects, which looks as if they were stored in multiple pools at the same time. We propose to store supertype and subtype information inside pools as references to the respective pools. The objects should be stored as a “double linked array list”³⁰, which is basically a linked list containing array lists:

Each pool stores the objects, with the static type of the pool in an array list. The pools of subtypes are stored in a linked list. Now an iterator over all instances of a type uses an iterator over all instances of a pool in combination with iterators over the pool and all its subpools. Therefore creating and deleting objects is an amortized $O(1)$ operation³¹ and it is guaranteed to maintain the semantic structure of the file, if IDs are not updated in phases other than reading and writing a file.

Note that the mere presence of distributed fields will change the worst case runtime complexity of these operations from (amortized) $O(1)$ to $O(\log(n))$, where n is the largest number of objects with a common distributed field in a state.

8.2.2 Reading Unmodified Data

Reading unmodified data is basically done by creating objects with ascending IDs while processing all eagerly processed fields. Pointer resolution in an unmodified state is an $O(t)$ operation, where t is the number of subtypes of the static type of the pointer³².

During the reconstruction of the initial dataset, an array in the base pool may be used to reduce the cost to $O(1)$ ³³. However, this helper array has to be dropped, as soon as the base pool is modified.

8.2.3 Modifying Data

The only legal way of modifying data is to access it through the generated API, which provides iterators, a type safe facade, factories and means of removing objects from states for each known type. A modification is any operation that might invalidate any existing object ID, i.e. deleting objects or inserting objects into non-empty storage

³⁰ In Java this would be a `LinkedList<ArrayList< ? extends B > >`, where `B` is the provided interface of the base type of a pool. Note however, that the `LinkedList`-part is expected to be realized using references between storage pools.

³¹ This can be achieved by appending new objects to the end of the pool and deleting existing objects by creating holes inside the pool. The compaction of the storage pool is done while writing the output and can be amortized into the writing costs.

³² This is usually a very small constant.

³³ The creation time for the array can be paid for during the creation of the base pool.

pools. Adding objects to empty storage pools does not count as a modification in the sense of a modified state, because it is not possible that a pointer to such an object lurks in an yet untreated field.

8.2.4 Reading Data in a Modified State

Reading Data from disk in a modified state, is very similar to reading data in the unmodified state. Except that resolution of stored pointers can no longer rely on the invariant that the ID of an object is also the index into the base type pool. There is a solution for this problem using $O(t + \log(n))$, where n is the number of instances with the same static type. However, the straightforward implementation is $O(t + n)$.

With this difference in mind, we strongly recommend adding a dirty flag to each storage pool which traces modifications. This is expected to eliminate the additional cost, because transformation of stored state is often the subject of monotonic growth, because each step of a computation usually adds instances of a type, which had no instances in a previous step.

8.2.5 Writing Data

Data which is lazy and modified can not directly be written to disk, because any data which can potentially refer to modified data has to be evaluated. After evaluating the respective data, serialization is straightforward. The evaluation of lazy data referring to potentially modified data can be done in $O(out)$, where out is the size of the output file. Writing the output is also in $O(out)$, thus writing is not per se a performance issue.

8.2.6 Final Thoughts on Runtime Complexity

Although the last sections read a lot like accessing serializable data is unnecessary expensive, this is in fact not the case.

Reading data without modifying is in $O(in)$, even if only a part of the data is read³⁴. This is mainly caused by the requirement of being able to process unknown data correctly. The actual cost should be limited by the cost of sequentially reading (or seeking through) the input file from disk.

Reading data, modifying it and writing it back is $O(in + m + out)$, which is not surprising at all, because one has to pay for reading the initial file, writing the complete output file and for performing the modifications.

Only the usage of a lot of non-monotonic lazy or distributed data is expensive. It is generally advised against using the lazy attribute if a field is read for sure during the lifetime of a serializable state. On the other hand, lazy fields can be very valuable, if their data is used for error reporting or debugging.

³⁴Note that a file can contain little payload compared to the type information stored in the file, therefore $O(in)$ is a sharp complexity estimate.

The problem is, or rather one of the problems, for there are many, a sizeable proportion of which are continually clogging up the civil, commercial, and criminal courts in all areas of the Galaxy, and especially, where possible, the more corrupt ones, this. The previous sentence makes sense. That is not the problem. This is: Change. Read it through again and you'll get it.

*Douglas Adams in So Long, and
Thanks for All the Fish*

Part III

Binary File Format

9 Impact of changing Specifications

If people are confronted with the specification of a file format, they tend to think that they know the format of that file. This is a fundamental problem and this section will briefly explain why this is not the case.

Let us think about a very simple tool description:

```
Tool {
    string name;
}
```

This description can be used by a tool chain driver to identify the tools that have been used to produce the content of a file. In fact this is useful information in a pre-internet world. In modern times, tools can be obtained over the internet, thus adding e.g. the location of a repository containing the tool can be used by a tool chain driver to run tools, that have not been installed on the target machine before. Thus, a modern feature may turn the specification into:

```
URL {
    ...
}
```

```
Tool {
    string name;
    /* not null, if a public repository containing the tool exists */
    URL repository;
}
```

Please note that changing the specification did not render old files incompatible to the new format, thus no data is lost! After some experiments with the new format, one might have observed, that the driver can execute tools which are not installed. This can be used to provide a file using a similar format that contains all known tools. The file can just be called `knownTools.sf` and can be processed by any tool that is able to read the tool info as described above.

The means to deal with format changes that are provided by SKILL are carefully designed. They are the result of examination of changes in the Bauhaus tool chain ,

An immediate consequence, besides many rules that deal with change, is that types and fields have names, which are only used to map them to expected types. For the sake of efficiency, those names are not used to link data together. Linkage is achieved by indices only.

10 Serialization

This section is about representing objects as a sequence of bytes. We will call this sequence *stream*, its formal type will be called *S*, the current stream will be called *s*.

To do (98)

To do (99)

To do (100)

To do (101)

To do (102)

To do (103)

We will assume that there is an implicit conversion between fixed sized integers³⁵ and streams. We also make use of a stream concatenation operator $\circ : S \times S \rightarrow S$.

We will use upper case letters for types (e.g. A) and lower case for instances of the respective type (e.g. a). The type \mathcal{T} denotes the set of types and τ a type. We will use τ_i for arbitrary type names and f_i for arbitrary fields, i.e. $\tau_i.f_j$ is the j 'th field of the i 'th type.

This section assumes that all objects about to be serialized are already known. It further assumes that their types and thus the values of the functions (i.e. `baseTypeName`, `typeName`, `index`, `[[_]]`) explained below can be easily computed.

The serialization function $[[_]]_\tau : \tau \times \mathcal{T} \rightarrow S$ maps an object $_$ of a type τ to a stream. Usually the type of the object can be inferred from context, thus we can simply write `[[_]]`. During the process of (de-)serialization, the type of an object can always be inferred from context.

To do (104)

To do (105)

To do (106)

10.1 Steps of the Serialization Process

In general it is assumed that the serialization process is split into the following steps:

1. All objects to be serialized are collected. This is usually done using the transitive closure of an initial set.
2. Objects are organized into their storage pools, i.e. the index function is calculated.
 - If the state was created by deserialization and indices have changed³⁶, fields using these indices have to be updated.
 - All known restrictions have to be checked.
3. The output stream is created as described below.

10.2 General File Layout

The file layout is optimized for fast appending of new objects. It is further optimized for lazy and partial loading of existing objects. It does also support type-safe and consistent treatment of unknown data structures. In order to achieve these goals, we have to store the type system used by the file together with the stored data. The type system itself is using strings for its representation. We want to be able to diagnose file corruption as early as possible, therefore most information stored in a file relies only on information that has already been processed. The only exception to this rule are field types referring to user types which are declared later in the same block.

Therefore the file is structured as an altering sequence of string blocks and type blocks, starting with a string block. The layout of string blocks (S) and type blocks (T) is visualized in figure 2, details will be explained in the following sub-sections.

To do (107)

To do (108)

To do (109)

To do (110)

To do (111)

To do (112)

To do (113)

³⁵As well as between fixed sized floating point numbers, because we define them to be IEEE-754 encoded 32-/64-bit sequences.

³⁶Indices can change by inserting objects before an existing object or by deleting objects. This is caused by the base pool index concept explained in section 10.3

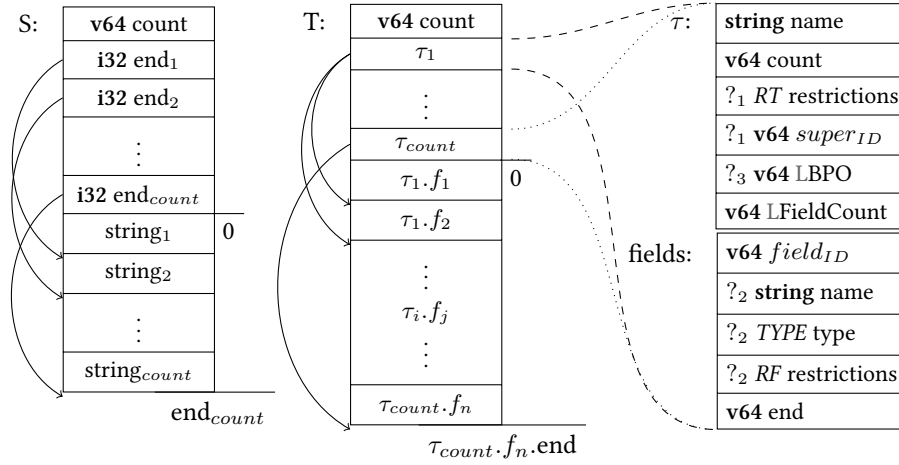


Figure 2: Visualization of the layout of string(S) and type(T) blocks. Type descriptors (τ) and their fields can drop fields in some contexts ($?_i$). Fields prefixed with a gray L contain information only relevant for the local block. The data chunks of blocks start at the respective 0 and reach until the respective end. Arrows indicate end-offsets – type blocks have one per field declaration. Field restrictions have to be placed after the fields type, because their translation may depend on the type.

10.2.1 Layout of a String Block

A string block starts with a v64 *count*, which stores the number of strings stored in the block. It is followed by *count* many i32 values, which store the offset³⁷ of the end of the respective string. The stored offsets split the data following after the last offset into utf-8 encoded strings.

The individual strings can be decoded using their index and the previous index (or 0, if there is no previous index). The strings stored in the string block are used by the following type blocks. For efficiency³⁸ reasons, strings used as type or field names shall be stored in lower case only. If a string containing user data equals a type/field name with different casing, a copy has to be made.

$$S(s_1, \dots, s_n) = \llbracket n \rrbracket_{v64} \llbracket end_1 \rrbracket_{i32} \dots \llbracket end_n \rrbracket_{i32} \llbracket s_1 \rrbracket_{utf8} \dots \llbracket s_n \rrbracket_{utf8}$$

To do (114)

10.2.2 Layout of a Type Block

Instances of types are organized into storage pools (see section 10.3), which are stored in type blocks.

A type block starts with a v64 *count*, which stores the number of *instantiated* types stored in the block. The *count* is followed by the respective amount of type declarations.

Type declarations themselves contain field declarations, which contain end-offsets into a data chunk located at the end of the type block, i.e. the field data is stored between the end-offset (or the start of the data chunk) of the previous field and the end-offset of this field. The Local Base Pool Start Index (LBPSI) field (marked with $?_3$ in Fig. 2) is only present if there is a super type and count is non-zero. The LBPSI

³⁷in Bytes

³⁸The lower case conversion during deserialization is optimized away.

is used to indicate which fields in a supertype declaration belong to instances of the current type. The local version of the Base Pool Start Index (BPSI) works in the same way as the BPSI in storage pools (see arrows in Fig. 5). The LBPSI field is only present if there is a supertype and the type has not yet been defined in a previous type block, because the data is not needed otherwise.

A type is *instantiated*, if the block adds new instances, fields or both. For example, the first block may add a type *node*, with an *ID* field. The second block, which is the result of a graph coloring tool, adds a *color* field to the *node*. We will come back to this example in section 10.2.3.

Type Order

TBD.

To do (115)

Type order is a partial order that linearizes type trees and ensures, that all data required for the interpretation of a type definition inside a type block is already present. This causes an improvement in both, error reporting and deserialization speed, as well as simplicity of the generated binding.

To do (116)

To do (117)

Effects of On Demand Deserialization

In this context, deferred deserialization means "the ability to skip data".

To do (118)

Inside of string blocks, most indices and string payload can be skipped. The offset type is `i32`, which allows for random access deserialization of individual strings, if the ID is known. This feature is very valuable, if there are many user data strings and few type names. The decision has also the consequence, that strings can only have 2^{32} bytes of data. We consider this limit irrelevant, because a user can still use a byte array to represents string-like objects exceeding this limit.

Inside of type blocks, all type information has to be processed in any case. The field data can be skipped completely.

Thus the laziest processing of a SKILL file will read the count fields of blocks, the last index of string blocks and all type information, including strings storing type and field names. Even in case of large files with many types, the amount of processed data is expected to be below a megabyte.

Effects of Appending

The desire to append new data of an arbitrary kind to an existing file, without having to rewrite existing data, affects mostly the hidden part of the generated language binding. From the perspective of the file format, appending adds altering sequence of blocks, instead of a single string block followed by a single type block. Means to add fields or instances to existing storage pools (see section 10.3) are made necessary by this feature as well. The omitted data is marked with $?_1$ and $?_2$ in Fig. 2: Fields marked with $?_1$ do only appear in the first block and are left away in all other blocks adding data of the respective type. Fields marked with $?_2$ are only present if the field is added to a type.

If appending is not used at all, the overhead, compared with a similar format that does not allow for appending, is about two bytes. If appending is used in a way, that only adds new fields or new storage pools, the overhead is still in the range of several bytes. Adding instances to existing storage pools is expected to create an overhead of about forty bytes, mostly caused by end-offsets of added field data.

If additional instances are added, the order of already instantiated fields is the order of their occurrence in previous blocks. If additional instances are added in combination with adding new fields, the new fields are located after the existing fields. The already existing fields contain data for the new instances, whereas new fields contain data for all existing instances.

Partial Fields

A field that lacks instance data or does not comply to its restrictions is called *partial*. If ownership is used, only tools owning the partial field, i.e. the surrounding type-tree, are allowed to access and modify the field. It is the duty of an owning tool to turn a partial field into a total field.

To do (119)

Fields with a default restriction will only be rendered partial if they do not comply to a restriction.

To do (120)

10.2.3 Node Example

Let us assume two tools with two SKILL specification files:

Listing 27: Specification of the Node Producer Tool

```
Node {
  i8 ID;
}
```

Listing 28: Specification of the Node Color Tool

```
Node {
  i8 ID;
  /** for the sake of simplicity a string like "red" */
  string color;
}
```

Let us assume, the first tool produces two nodes ("23" and "42") and the second appends color fields to these nodes ("red" and "black"). The layout of the resulting file, after the second tool is done, is given in figure 3. Actual field data is kept abstract, because serialization of field data will be explained below. The data produced by the first tool is S and T, the second tool produces S' and T'.

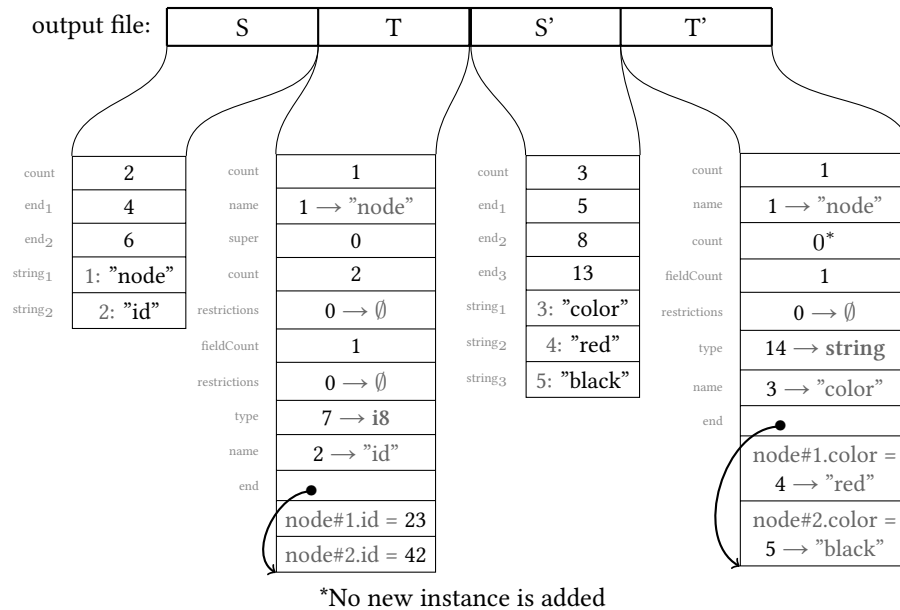


Figure 3: Illustration of the file obtained after running the example tool chain. Gray parts are the interpretations of the respective value and will be explained below. Light gray annotations shall serve as a reminder of the meaning of the contents of the respective field. In section 10.6 there is a more complete example.

We will see in the next section, that fields referring to other objects are stored by reference. Strings are somewhat a special case of this. Strings are used both, as a first

class type and to represent type information itself. This has the consequence, that string pools contain user data and data, which might not be directly observed by a user.

Adding instances

Now let us assume, that the first tool is ran twice, adding ("23" and "42") in the first run and ("-1" and "2") in the second; we wont run the coloring tool, thus S/T are the result of the first run and S*/T* are the result of the second run.

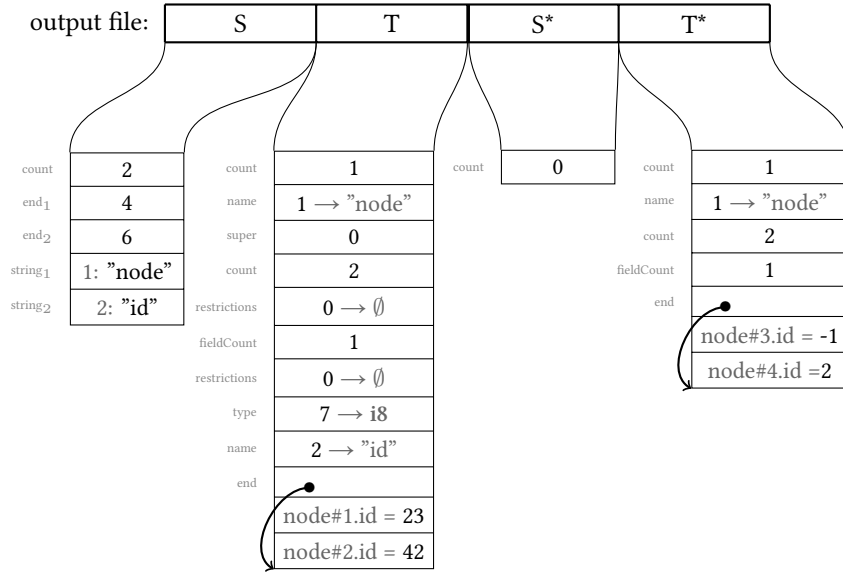


Figure 4: Illustration of the file obtained after running the example tool chain. Gray parts are the interpretations of the respective value and will be explained below. Light gray annotations provide contents descriptions. Note that string blocks (and type blocks) can be empty, but must not be omitted.

10.3 Storage Pools

This section contains the serialization function for an individual storage pool. A storage pool stores the instances of a type, i.e. all field data of the type of the pool. We assume that storage pools are not empty. The serialization of an empty storage pool is its name and super information, with all other values set to zero. This behavior ensures the integrity of the written file while minimizing size at the same time.

Writing objects of a pool requires the following functions: $baseTypeName : \mathcal{U} \rightarrow S$, $typeName : \mathcal{U} \rightarrow S$ and $index : \mathcal{U} \cup \{\mathbf{string}\} \rightarrow S$.

The $baseTypeName$ is either the index of the string used as type name of the base type or 0x00. The $typeName$ is the index of the string used as type name of the argument type. The index is the unique index of the argument object. Indices are assigned ascending from 1 for instances of a base type, for each base type. For example the index 23 can be given to a *Node* object and a string "hello", because they do not

To do (121)

To do (122)

have a common base type. Indices do not have holes. Although indices are serialized using the v64 type, they are treated as if they were unsigned integers³⁹

A basic concept of the serialization format is to store the data grouped by type into storage pools. This concept enables us to obtain type information of objects from their position in a file, instead of storing a type descriptor with each object.

If objects are referred to from other objects, those references are given as an integer, which is interpreted as index into the respective storage pool. The NULL pointer is represented by the index 0.

Each pool knows how many instances it holds. Storage pools with a supertype store the name of the supertype and a BPSI. Further, we assume that objects, which are written to a file, have indices such that for any type, all instances of that type have adjacent indices.

A short example (Fig. 5) illustrates the concept. It contains five types A, B, C, D, and N. For the sake of simplicity, each type has a single field of an arbitrary type τ_x (serialization of field data will be explained in the next section). A and N are base types. A has 6 instances. B, C are subtypes of A; D is a subtype of B. B/C/D have 4/1/1 instances and BPSIs of 2/6/5, respectively. The arrows represent the end-offsets stored in the field descriptions, which are used to separate the data part of the type block into field data. The *index*-row displays the index of the object that belongs to the serialized field.

To do (123)

For the sake of readability, the header of the type pool is omitted in the stream part of the picture.

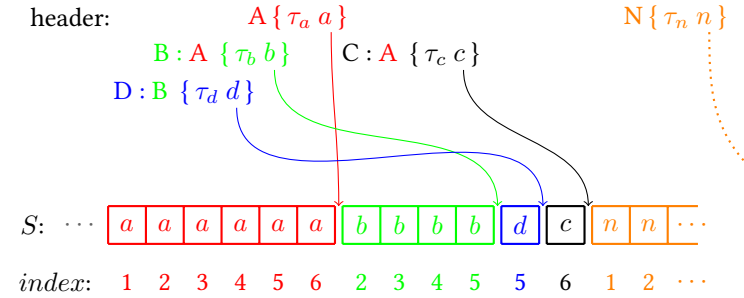


Figure 5: Field data of several pools stored in a type block. Arrows indicate the values of data fields of the respective field description in the header of the type block. *S* shows a part of the data chunk of a type block.

Although illustrations above might suggest an order of field data, this is in fact not the case. This allows for optimization during serialization, such as multi-threaded encoding of different fields at the same time.

10.3.1 Subtypes

TBW

To do (124)

10.3.2 Appending Fields to existing Pools

Appending a field to an existing pool is rather easy, because existing objects are not directly modified. The respective type block will have a type declaration with a field

³⁹That is the index -1 refers to object number $2^{64} - 1$.

declaration of the added field. The added field data is stored in the data part of the type block.

The order in which existing fields are appended to a file is unconstrained, because all required information can be obtained by an inexpensive lookup.

10.3.3 Appending Objects to existing Pools

To do (125)

If objects are added to existing pools, all type pools (i.e. the pools of the type and all its supertypes) have to be updated. If a supertype exists, the local base pool start index will give the start index of the added objects in the local pool. With the types as in figure 5 we can for example add several object in three blocks.

Let T1 contain 6 objects of types aabbbc, T2 contain 4 objects of types bbdd and T3 another 3 objects acd.

T1 contains the full type information of types A,B and C. There are 6 A instances, 3 B instances (with LBPSI=3) and 1 C instance (with LBPSI=6).

T2 contains only the type name of A and B (LBPSI=1) and field data for 4 instances each. Additionally, there is a new type D (with LBPSI=3) and D-field data for two instances.

T3 contains field data for A, B (LBPSI=3), C (LBPSI=2) and D (LBPSI=3) objects.

The in-memory representation after loading these objects is expected⁴⁰ to be either aabbbcbddacd or aaabbbbbbddcc, depending on used hints and restrictions .

To do (126)

10.4 Serialization of Field Data

In this section, we want to describe the serialization of individual fields using the function $\llbracket _ \rrbracket_\tau$. The serialization of an object is done by serializing all its fields into the stream. In this section, we assume that the three functions defined in the last section are implicitly converted to streams using the v64 encoding. We assume further, that compound types provide a function $size : \mathcal{T} \rightarrow \mathcal{I}$, which returns the number of elements stored in a given field. Let f be a non-constant⁴¹ non-auto field of type t , then $\llbracket f \rrbracket_t$ is defined as⁴²

- $\forall t \in \mathcal{U} \cup \{\text{string}\}. \llbracket f \rrbracket_t = \begin{cases} 0x00, & f = \text{NULL} \\ \llbracket index(f) \rrbracket_{v64} & \text{else} \end{cases}$
- $\llbracket f \rrbracket_{\text{annotation}} = \begin{cases} 0x00 \ 0x00, & f = \text{NULL} \\ \llbracket baseTypeName(f) \rrbracket_{v64} \circ \llbracket index(t) \rrbracket_{v64} & \text{else} \end{cases} \quad 43$
- $\llbracket \top \rrbracket_{\text{bool}} = \llbracket \varepsilon x.x \neq 0 \rrbracket_{i8} \quad 44$
- $\llbracket \perp \rrbracket_{\text{bool}} = 0x00$
- $\forall t \in \mathcal{I} \setminus \{\mathbf{v64}\}. \llbracket f \rrbracket_t = f$
- $\llbracket f \rrbracket_{\mathbf{v64}} = encode(f) \quad 45$

⁴⁰ The in-memory representation is intentionally left unspecified. The expected behavior will be the result of a straightforward implementation.

⁴¹Constant fields are not serialized, because their value is already stored in the type declaration.

⁴²We will use C-Style hexadecimal integer literals for integers in streams.

⁴³We do not want to use type IDs here, because we do not want to touch all annotation fields if we modify the type Pool.

⁴⁴The usual solution is either 0x01 or 0xFF.

⁴⁵With encode as defined in listing C.1.

- $\llbracket f \rrbracket_{f32} = \llbracket f \rrbracket_{f64} = f^{46}$
- $\forall g \in \mathcal{B}, n \in \mathbb{N}^+. t = g[n] \implies \llbracket f \rrbracket_t = \bigcirc_{i=0}^{n-1} \llbracket f_i \rrbracket_g$
- $\forall g \in \mathcal{B}, n = \text{size}(f), t \in \{g[], \text{set}\langle g \rangle, \text{list}\langle g \rangle\}. \llbracket f \rrbracket_t = \llbracket n \rrbracket_{v64} \bigcirc_{i=0}^{n-1} \llbracket f_i \rrbracket_g$
- $\forall r, s, t \in \mathcal{T}. \llbracket _ \rrbracket_{\text{map}\langle r, s, t \rangle} = \llbracket _ \rrbracket_{\text{map}\langle r, \text{map}\langle s, t \rangle \rangle}$
- $\forall k, v \in \mathcal{T}, n = \text{size}(f), t = \text{map}\langle k, v \rangle. \llbracket f \rrbracket_t = \llbracket n \rrbracket_{v64} \bigcirc_{i=0}^{n-1} \llbracket f.k_i \rrbracket_k \llbracket f[k_i] \rrbracket_v$
- $\llbracket rs \rrbracket_{RT} / \llbracket rs \rrbracket_{RF} = \text{see section 7.1.}$

Note that each restriction has its own serialization function. Common properties are specified and explained at the very beginning of section 7.1.

$$\bullet \llbracket t \rrbracket_{\text{type}} = \begin{cases} \llbracket id \rrbracket_{v64} \circ \llbracket val \rrbracket_t & id \in [0, 4] \\ \llbracket id \rrbracket_{v64} & id \in [5, 14] \\ 0x0F \circ \llbracket i \rrbracket_{v64} \circ \llbracket T \rrbracket_{\text{type}} & t = T[i] \\ 0x11 \circ \llbracket T \rrbracket_{\text{type}} & t = T[] \\ 0x12 \circ \llbracket T \rrbracket_{\text{type}} & t = \text{list}\langle T \rangle \\ 0x13 \circ \llbracket T \rrbracket_{\text{type}} & t = \text{set}\langle T \rangle \\ 0x14 \circ \llbracket K \rrbracket_{\text{type}} \circ \llbracket V \rrbracket_{\text{type}} & t = \text{map}\langle K, V \rangle \\ \llbracket 32 + \text{storagePoolIndex}(t) \rrbracket_{v64} & t \in \mathcal{U} \end{cases}$$

ids of restrictions and types are listed in appendix G. The holes are intentional and enable future built-in types without breaking the file format.

To do (127)

Note that the function *storagePoolIndex* is implicitly given by the order of storage pools appearing in a file and has therefore to be computed both upon serialization and deserialization. The first *storagePoolIndex* is 0.

To do (128)

10.5 Endianness

Files are stored in network byte order, as described in RFC1700, Page 2 [RP94].

If a client is running on a little endian machine, the endianness has to be corrected, both when reading and writing files. This can be achieved by changing the implementation of $\llbracket _ \rrbracket_{i*}$ and $\llbracket _ \rrbracket_{f*}$ -translations. Note that some standard libraries provide functions to read and write binary data in network byte order.

10.6 Date Example

This section will provide a concise example of how serialization takes place. For the sake of simplicity and brevity, we will serialize two objects of a single (simple) type into a stream. We will use the following file format:

To do (129)

Listing 29: Date File Format

```
Date {
    v64 date ;
}
```

⁴⁶Assuming the float to be IEEE-754 encoded (see [iee08] §3.4), which allows for an implicit bit-wise conversion to fixed sized integer.

Further, we want to store two objects with *date* values 1 and -1.

To do (130)

The first thing we have to do is to collect the objects to be stored. In a set-theory inspired notation with square brackets indicating records we would get something like:

$$Date = \{[date : 1], [date : -1]\}$$

Now we have to create storage pools. We start with the creation of type information. The date storage pool looks something like this:

$$[name : "date", super : 0, count : 2, restr. : \emptyset, fields[\\ 1 : [restr. : \emptyset, t : v64, n : "date", [1, -1]]]$$

To do (131)

Now we can see, that we have a string, which has to be serialized, but is not yet in the string pool (which was empty up until now), so we create a string pool⁴⁷ as well:

To do (132)

$$[1 : "date"]$$

Now we can start to encode the pools:

To do (133)

1. write the string block:

$$\begin{aligned} \llbracket [1 : "date"] \rrbracket &= \llbracket 1 \rrbracket_{v64} \circ \text{offset}("date") \circ \llbracket "date" \rrbracket \\ &= \llbracket 1 \rrbracket_{v64} \circ \llbracket 4 \rrbracket_{i32} \circ 64 \ 61 \ 74 \ 65 \\ &= 01 \ 00 \ 00 \ 00 \ 04 \ 64 \ 61 \ 74 \ 65 \end{aligned}$$

2. write the type block:

To keep things readable, we will first encode everything but the field data:

$$\begin{aligned} &\llbracket [name : "date", super : 0, count : 2, restr. : \emptyset, \dots] \rrbracket \\ &= \llbracket 1 \rrbracket_{v64} \circ \llbracket "date" \rrbracket \circ 0 \circ \llbracket 2 \rrbracket_{v64} \circ \llbracket \emptyset \rrbracket_R \\ &= 1 \circ \text{index}("date") \circ 0 \circ 2 \circ \llbracket \text{size}(\emptyset) \rrbracket_{v64} \\ &= 1 \circ 1 \circ 0 \circ 2 \circ 0 = 01 \ 01 \ 00 \ 02 \ 00 \end{aligned}$$

Now we can continue with the field:

To do (134)

$$\begin{aligned} &\llbracket [\dots fields[1 : [restr. : \emptyset, t : v64, n : "date", [1, -1]]]] \rrbracket \\ &= \llbracket \text{size}(fields[\dots]) \rrbracket \circ \llbracket [restr. : \emptyset, t : v64, n : "date", \text{offset}([1, -1])] \rrbracket \circ \\ &\llbracket [1, -1] \rrbracket \\ &= 1 \circ \llbracket \emptyset \rrbracket_{rest} \circ \llbracket v64 \rrbracket_{type} \circ \llbracket "date" \rrbracket \circ \text{offset}("date.date") \circ \llbracket [1, -1] \rrbracket \\ &= 1 \circ 0 \circ 0B \circ \text{index}("date") \circ \text{offset}("date.date") \circ \llbracket 1 \rrbracket_{v64} \circ \llbracket -1 \rrbracket_{v64} \\ &= 1 \circ 0 \circ 0B \circ 1 \circ \text{offset}("date.date") \circ 01 \circ FF \ FF \ FF \ FF \ FF \ FF \ FF \ FF \ FF \ FF \\ &= 1 \circ 0 \circ 0B \circ 1 \circ \llbracket 10 \rrbracket \circ 01 \circ FF \ FF \ FF \ FF \ FF \ FF \ FF \ FF \ FF \ FF \\ &= 01 \ 00 \ 0B \ 01 \circ 0A \circ 01 \ FF \ FF \ FF \ FF \ FF \ FF \ FF \ FF \ FF \ FF \\ &= 01 \ 00 \ 0B \ 01 \ 0A \ 01 \ FF \ FF \ FF \ FF \ FF \ FF \ FF \ FF \ FF \ FF \end{aligned}$$

The type block is now serialized to the stream:

$$01 \ 01 \ 00 \ 02 \ 00 \ 01 \ 00 \ 0B \ 01 \ 0A \ 01 \ FF \ FF \ FF \ FF \ FF \ FF \ FF \ FF \ FF \ FF$$

3. writing the output:

The remaining work is just to write the string block and the type block to a file, starting with the string block, so we get:

$$\begin{aligned} &01 \ 00 \ 00 \ 00 \ 04 \ 64 \ 61 \ 74 \ 65 \ 01 \ 01 \ 00 \ 02 \ 00 \ 01 \ 00 \\ &0B \ 01 \ 0A \ 01 \ FF \ FF \ FF \ FF \ FF \ FF \ FF \ FF \ FF \ FF \end{aligned}$$

Deserialization of this stream is explained in section 11.2.

⁴⁷The string pool, unlike regular storage pools is just a length encoded array of strings.

10.7 Map Example

In this short section will give an example on how to serialize a single map field. For the sake of brevity and simplicity, we will encode a map of type `map<i8, i8, i8>`. The maps data will be $m = (-1 \rightarrow (-2 \rightarrow -3, -3 \rightarrow -3), -2 \rightarrow (-1 \rightarrow -2))$. The entries are all negative, in order to make distinction between coded sizes and values obvious.

The coding of a map is basically a recursion on the remaining map types, consuming types from left to right. This leads to two cases:

1. The remaining map is at least binary:

We have to store the number of keys and for each key, we have to store the key, followed by the serialization of the referred value. Remember that `map<U, V, T>` is right associative, i.e. it is treated like `map<U, map<V, T>>`.

2. There is only one type left:

This means, that we have to encode the value of the last map and we are done.

Serialization

```

[[m]]
  by rule 1:
= [[2]] ∘ [[-1]] ∘ [[m(-1)]] ∘ [[-2]] ∘ [[m(-2)]]
= 02 FF ∘ [[m(-1)]] ∘ [[-2]] ∘ [[m(-2)]]
  by rule 1 on m(-1):
= 02 FF ∘ [[2]] ∘ [[-2]] ∘ [[m(-1)(-2)]] ∘ [[-3]] ∘ [[m(-1)(-3)]] ∘ [[-2]] ∘ [[m(-2)]]
= 02 FF 02 FE ∘ [[m(-1)(-2)]] ∘ [[-3]] ∘ [[m(-1)(-3)]] ∘ [[-2]] ∘ [[m(-2)]]
  by rule 2 on m(-1)(-2&-3):
= 02 FF 02 FE ∘ [[-3]] ∘ [[-3]] ∘ [[-2]] ∘ [[m(-2)]]
= 02 FF 02 FE FD FD FD FE ∘ [[m(-2)]]
  by rule 1 on m(-2):
= 02 FF 02 FE FD FD FD FE ∘ [[1]] ∘ [[-1]] ∘ [[m(-2)(-1)]]
= 02 FF 02 FE FD FD FD FE 01 FF ∘ [[m(-2)(-1)]]
  by rule 2:
= 02 FF 02 FE FD FD FD FE 01 FF ∘ [[-2]]
= 02 FF 02 FE FD FD FD FE 01 FF FE

```

Colored Data

```

(-1 → (-2→-3, -3→-3), -2 → (-1→-3))
02 FF 02 FE FD FD FD FE 01 FF FE

```

11 Deserialization

Deserialization is mostly straightforward.

To do (135)

The only notable property is that a binding has to ensure that no objects can be allocated that have more fields in the deserialized file than in the specification. This is important because it prevents from violating invariants of other tools in a tool chain by allocating objects through a partial view onto a type.

The general strategy is:

1. the string block is processed

2. the header of the type block is processed
3. required fields are parsed using the type and position information obtained from the respective block
4. until the end of file has been reached, goto step 1

11.1 An LL-inspired Approach to sf-Parsing

To do (136)

Various approaches to parsing have shown, that the most successful and straightforward approach to parsing a binary SKILL file is an LL(1)-style recursive descent parser. The format is designed, such that context sensitive parsing can be done easily using known parts of the state.

In this section, we will sketch a lexer-parser-implementation using notation and specification from section 10.

11.1.1 Lexer

The lexer provides the parser with tokens for

- bool
- i8, ..., i64
- v64
- f32, f64
- utf8 character arrays, as they occur in string blocks

Furthermore, the lexer shall provide an EOF token allowing the parser to stop after a type block. The EOF token can be implemented implicitly by adding a query to the lexer. For the sake of efficiency, the lexer should provide means of random access to a file.

Errors The lexer can only fail, if the parser requests a token beyond the end of file.

11.1.2 Parser State

We assume that the parser provides an implementation for the functions

- *string(i)*: The *i*'th string stored in the current file.
- *type(i)*: The *i*'th type definition stored in the current file.
- *type(n)*: The type definition of the name *n*, as obtained from the current file.
- *super(t)*: The super type of *t* or \perp if none exists.
- *fields(t)*: The fields defined for type *t*.

We assume further, that there is are *string.next* and *type.next* functions, that add a new string or type to the state.

To do (137)

Note String and type are basically arrays, while other queries should be fields of a class or record structure.

11.1.3 Parser

We use the operator "→" (read as "is") to store the value of a token or rule in a variable. We use indentation to indicate the scope of a loop. Parsing starts with the rule `File`.

File Files consist of string and type blocks:

```
while (!EOF)
  StringBlock
  TypeBlock
```

In order to improve error reporting, the number of already processed blocks should be stored and included in error messages.

StringBlock Using local variables `length` and `offsets` we get:

```
v64 → length
for (0 ≤ i ≤ length)
  v64 → offsets(i)

for (0 ≤ i ≤ length)
  utf8[offset(i)] → string.next
```

The string block should be handled lazily. This can be achieved by storing the offsets in the parsers state accessing the strings on demand.

TypeBlock Using the local variables `length` and `newFields` we get:

```
v64 → length
for (0 ≤ i ≤ length)
  TypeDefinition → (type.next, newFields.next)

for (f in newFields)
  FieldData
```

The `FieldData` rule is similar to the definitions in section 10.4.

TypeDefinition A heavily context-sensitive rule:)

```
string(v64) → name
if ( ∃ type(name))
  ..
else
  ..
```

The `∃type` checks, whether we defined the type in a previous block. If the type has been defined in this block, an error has to be reported.

11.2 Date Example

To do (138)

Let d be the deserialization function – basically the inverse function of $[_]$. We want to read the sequence we created during the serialization example in section 10.6:

```
01 00 00 00 04 64 61 74 65 01 01 00 02 00 01 00
0B 01 0A 01 FF FF FF FF FF FF FF FF FF
```

1. deserialization of a string block:

$d(010000000464617465010100 \dots)$
→ a string block starts with a v64 indicating the number of strings stored inside
 $d(01)d(0000000464617465010100 \dots)$
→ we got one string
 $d(01)d(00000004)d(64617465010100 \dots)$
→ the next string has 4 bytes, the block ends in 4 bytes
 $string[1 : d(64617465)]d(0101 \dots)$
→ build the string pool with the first string block
 $string[1 : "date"]d(0101 \dots)$
→ we processed the string directly, because lazy evaluation makes the example rather confusing

2. deserialization of a type block (reading the header):

$d(01)d(0100020001000B010A01 \dots)$
→ there is one type definition in this block; read its name
 $d(01)d(00020001000B010A01 \dots)$
 $= "date"d(00)020001000B010A01 \dots)$
→ we do not know the type "date" yet (in terms of processing the file), so we expect super type information, count, restrictions and field declarations
 $[name : "date", super : 0]d(02)d(00)d(01)d(000B010A01FF \dots)$
→ Date has no super type, thus the next field is not a local BPSI and we can read until the number of fields.
 $[name : "date", super : 0, count : 2, restr. : \emptyset, fields[1 : _]]$
 $d(00)d(0B)d(01)d(0A)01FF \dots)$
→ We did not know date yet, thus we do not know the type of the first field. Therefore the next information are the fields restrictions, type, name and offset.
 $[name : "date", super : 0, count : 2, restr. : \emptyset, fields[1 : [restr. : \emptyset, t : v64, n : "date", offset : 10]]]d(01FF \dots)$
→ We read all data belonging to the header of this type block, thus we know that the next 10 bytes are field data. The file ends after the next 10 Bytes, so we can continue constructing objects.

3. construction of date objects:

$Date = \{[date : _], [date : _]\}d(01FF \dots)$
→ using the data from the header, we created two date objects; the only thing left is parsing field data for the date fields and setting the data accordingly
 $Date = \{[date : 1], [date : _]\}d(FF \dots)$
→ the first field is a 1

$Date = \{[date : 1], [date : -1]\}$

→the second field is a -1; we reached the end of the field data on the last object, so everything is fine – no additional or missing data

Unsurprisingly, the restored objects are exactly the objects we serialized in section 10.6.

To do (139)

There is a theory which states that if ever anyone discovers exactly what the Universe is for and why it is here, it will instantly disappear and be replaced by something even more bizarre and inexplicable. There is another theory which states that this has already happened.

Douglas Adams in The Restaurant at the End of the Universe

Part IV

Future Work

Further research has to be done in the area of restrictions. Especially the effect of mixins on the usage of skill can not be predicted at this moment. To do (140)

A general purpose viewer and editor for SKill files should be implemented.

We will look into ways of implementing type-safe unions. It has to be evaluated, whether they should be first class citizens of the SKill type system, or if they can be represented with interfaces and hints or restrictions. We will also evaluate possible ways to support type casts, in order to allow for using language specific types, such as bit-fields.

We will introduce a general assertion restriction, which can be used to assert per instance invariants with sufficiently powerful expressions.

We will evaluate a template import statement, which can be used to import files together with a substitution. That way, one can create more complex data structures such as B-Trees. This feature is not a priority and only useful for large files and projects and requires some kind of type casts to yield the desired effect automatically.

We will further evaluate means of supporting common high level specification tasks, such as creating effective views for individual tools of a tool chain. This can either be in the form of a skill specification editor or an extended specification language.

Although there have been experiments with name spaces, the mechanism has been deferred to future versions of SKill. This is the case, because there was no obvious easy-to-use implementation for existing language bindings. Furthermore many questions regarding name spaces remained open, such as

- Can name spaces have comments?
- Can they have hints or restrictions?
- What if multiple uses of the same name space have contradicting modifiers?
- Share name spaces a name space with type definitions?

Numerical limits, as described in appendix F, will be dropped as soon as the top 20 programming languages support 64bit index types. To do (141)

Part V

Appendix

A Full Grammar of the Specification Language

$\langle file \rangle$::= $\langle header \rangle \langle definitions \rangle$
$\langle header \rangle$::= $\langle head-comment \rangle^* \langle include \rangle^*$
$\langle head-comment \rangle$::= $'\# (\sim[\backslash n])^* \backslash n'$
$\langle include \rangle$::= $(include with) \langle string \rangle^+$
$\langle declaration \rangle$::= $\langle namespace \rangle$ $\langle user-type \rangle$ $\langle enum-type \rangle$ $\langle interface-type \rangle$ $\langle field-change \rangle$
$\langle namespace \rangle$::= $\langle description \rangle \text{ namespace } \langle ID \rangle \{' \langle declaration \rangle^+ \}'$
$\langle user-type \rangle$::= $\langle change-modifier \rangle? \langle description \rangle \langle ID \rangle ((: with extends) \langle ID \rangle)^* \{' \langle field \rangle^* \}'$
$\langle change-modifier \rangle$::= $++ \mid -- \mid ==$
$\langle enum-type \rangle$::= $\langle comment \rangle? \text{ enum } \langle ID \rangle \{' \langle ID \rangle (, \langle ID \rangle)^* ;' \langle field \rangle^* \}'$
$\langle interface-type \rangle$::= $\langle comment \rangle? \text{ interface } \langle ID \rangle ((: with extends) \langle ID \rangle)^* \{' \langle field \rangle^* \}'$
$\langle field-change \rangle$::= $\langle change-modifier \rangle \langle description \rangle (\text{auto})? \langle type \rangle \langle ID \rangle \{' \langle ID \rangle ;' \}$
$\langle field \rangle$::= $\langle description \rangle ((\langle constant \rangle) (\langle data \rangle)) ;'$
$\langle constant \rangle$::= $\text{const } \langle type \rangle \langle ID \rangle '=' \langle int \rangle$
$\langle data \rangle$::= $(\text{auto})? \langle type \rangle \langle ID \rangle$

Table 1: SKILL Specification

$\langle type \rangle$	$::= \text{map } \langle type\text{-multi} \rangle$ $\text{set } \langle type\text{-single} \rangle$ $\text{list } \langle type\text{-single} \rangle$ $\langle array\text{-type} \rangle$
$\langle type\text{-multi} \rangle$	$::= ' < ' \langle ground\text{-type} \rangle (' , ' \langle ground\text{-type} \rangle)_+ ' > '$
$\langle type\text{-single} \rangle$	$::= ' < ' \langle ground\text{-type} \rangle ' > '$
$\langle array\text{-type} \rangle$	$::= \langle ground\text{-type} \rangle ('[' \langle INT \rangle ? '] ') ?$
$\langle ground\text{-type} \rangle$	$::= \text{annotation} \mid \langle ID \rangle$

Table 2: SKill Specification Types

$\langle description \rangle$	$::= \langle comment \rangle ? (\langle restriction \rangle \langle hint \rangle)^*$
$\langle comment \rangle$	$::= /* \langle comment\text{-text} \rangle (\langle comment\text{-TAG} \rangle : ? \langle comment\text{-text} \rangle)^* */$
$\langle comment\text{-text} \rangle$	$::= (\langle comment\text{-prefix} \rangle (\sim [* / \langle comment\text{-TAG} \rangle])^*)^*$
$\langle comment\text{-prefix} \rangle$	$::= '\backslash n' \langle whitespace \rangle ('*' \langle whitespace \rangle) ?$
$\langle comment\text{-TAG} \rangle$	$::= '@' [a-z]^+$
$\langle restriction \rangle$	$::= '@' \langle ID \rangle ('(' (\langle r\text{-arg} \rangle (',' \langle r\text{-arg} \rangle)^*) ?) ') ?$
$\langle r\text{-arg} \rangle$	$::= \langle FLOAT \rangle \mid \langle INT \rangle \mid \langle STRING \rangle \mid (\langle ID \rangle ('?' \langle ID \rangle)^*)$
$\langle hint \rangle$	$::= '! ' \langle ID \rangle$

Table 3: SKill Specification Descriptions

B Full Grammar of the Binary File Format

The notation used to specify the file format is explained in chapter 10. Additionally we use small fracture letters for blocks, to avoid collisions. This is basically a synopsis of chapters 10 and 7.1.

$$\begin{aligned}
File &= s_1 \circ t_1 \circ \dots \circ s_n \circ t_n \\
s &= [n]_{v64} \circ [end_1]_{i32} \dots [end_n]_{i32} \circ [string_1]_{utf8} \dots [string_n]_{utf8} \\
t &= [n]_{v64} \circ [t_1]_T \dots [t_n]_T \circ [t_1.f_1] \dots [t_i.f_j] \dots [t_n.f_m] \\
[t]_T &= [name]_{string} \circ [t]_{?1} \circ [count]_{v64} \circ [\#fs]_{v64} \circ [field_1]_{\mathbb{F}} \dots [field_{\#fs}]_{\mathbb{F}} \\
[t]_{?1} &= \begin{cases} [rs]_{RT} \circ [t]_{?3} & \text{first occurrence of } t \text{ in } S \\ \varepsilon & \text{else} \end{cases}
\end{aligned}$$

$$\begin{aligned}
\llbracket t \rrbracket_{?3} &= \begin{cases} \llbracket super.typeID - 31 \rrbracket_{v64} \circ \llbracket LBPO \rrbracket_{v64} & \exists super \\ \varepsilon & else \end{cases} \\
\llbracket f \rrbracket_{\mathbb{F}} &= \llbracket ID \rrbracket_{v64} \circ \llbracket f \rrbracket_{?2} \circ \llbracket end \rrbracket_{v64} \\
\llbracket t \rrbracket_{?2} &= \begin{cases} \llbracket name \rrbracket_{string} \circ \llbracket rs \rrbracket_{RF} \circ \llbracket t \rrbracket_{TYPE} & \text{first occurrence of } f \text{ in } S \\ \varepsilon & else \end{cases}
\end{aligned}$$

To do (142)

C Variable Length Coding

Size and Length information is stored as v64, i.e. variable length coded 64 bit unsigned integers (aka C's `uint64_t`). The basic idea is to use up to 9 bytes, where any byte starts with a 1 iff there is a consecutive byte. This leaves a payload of 7 bit for the first 8 bytes and 8 bits of payload for the ninth byte. This is very similar to the famous utf8 encoding and is motivated, as it is the case with utf8, by the assumption, that smaller numbers are a lot more likely. It has the nice property, that there are virtually no numerical size limitations. The following small C++ functions will illustrate the algorithm:

To do (143)

```

uint8_t* encode(uint64_t value) {
    // calculate effective size
    int size = 0;
    {
        uint64_t buckets = value;
        while(buckets) {
            buckets >>= 7;
            size++;
        }
    }
    if(!size) {
        uint8_t[] result = new uint8_t[1];
        result[0] = 0;
        return result;
    } else if(10==size)
        size = 9;

    // split
    uint8_t[] result = new uint8_t[size];
    int count=0;
    for(; count < 8 && count < size-1; count++) {
        result[count] = value >> (7*count);
        result[count] |= 0x80;
    }
    result[count] = value >> (7*count);
    return result;
}

```

Note that the encode function can be heavily simplified if a write buffer with a capacity of at least 9 Bytes is available.

Listing C.1: Variable Length Encoding

```

uint64_t decode(uint8_t* v64) {
    int count = 0;
    uint64_t result = 0;
    register uint64_t bucket;
    for(; count < 8 && (*v64)&0x80; count++, v64++) {
        bucket = v64[0];
        result |= (bucket&0x7f) << (7*count);
    }
    bucket = v64[0];
    result |= (8==count?bucket:(bucket&0x7f)) << (7*count);
    return result;
}

```

Note that the decode function can be improved using manual loop unrolling in most programming language compiler combinations.

Listing C.2: Variable Length Decoding

Optimized versions can be found in the Scala Generator. See .

To do (144)

To do (145)

D Error Reporting

This section describes some errors regarding ill-formatted files, which must be detected and reported. The order is based on the expected order of checking for the described error. The described errors are expected to be the result of file corruption, format change or bugs in a language binding.

To do (146)

To do (147)

To do (148)

To do (149)

Serialization

- TBD

Deserialization

- Type names have to be checked for illegal characters. This is very important in order to detect future format extensions. Such a file has to be rejected.
- If EOF is encountered unexpectedly, an error must be reported before producing any observable result.
- If an index into a pool is invalid⁴⁸, an error must be reported.
- If the deserialization of a storage pool does not consume exactly the `sizeBytes` specified in its header, an error must be reported. This is a strong indicator for a format change.
- If the serialized type information contains cycles, an error has to be reported, which contains at least all type names in the detected cycle and the base type, if one can be determined.

To do (150)

To do (151)

⁴⁸because it is larger then the last index/size of the pool

- If a storage pool contains instances which, based on their location⁴⁹ in the base pool, should be subtypes of some kind, but have no respective subtype storage pool, an error must be reported with at least, the base type name, the most exact known type name and the adjacent base type names. This is a strong indicator for either file corruption or a bug in the previously used back-end.
- All known constant fields of a type have to be checked before producing any observable objects of the respective type. If some constant value differs from the expected value, an error must be reported, which contains at least the type, the field type and name, the type block index, the total number of type blocks, the expected value and the actual value.
- If a serialized value violates a restriction or the invariant of a type,⁵⁰ an error must be reported as soon as this fact can be observed⁵¹.

E Levels of Language Support

The implementation of a fully tested SKiL binding is a time consuming and error prone process. In order to keep things simple, the language has been divided into three parts. The basic level can be implemented in a short period of time and suffices for many SKiL related tasks. The core level contains all common SKiL functionality. Core level bindings can usually be created by extending a basic level binding. The full level contains all SKiL features. Creating and testing an efficient full featured binding is expected to be time consuming and not relevant for many use cases. Experience shows that many designs for core level bindings can not be extended to full featured SKiL bindings. The Architecture of a full SKiL binding is explained in chapter 8.

To do (152)

As a consequence, any binding may be called SKiL binding that implements the core language features. The list below shall also serve as a means of talking about the abilities provided by a specific language binding (generator).

- Basic
 - Integer types i8 to i64 and v64
 - `string` and `bool`
 - reading of files exactly matching the specification
 - writing of files exactly matching the specification
 - means of SKiL state management
 - User types with sub-typing
- Core
 - `annotation`
 - Floats
 - Compound types
 - Bitvector

To do (153)

⁴⁹I.e. after a subtype but not inside another subtype

⁵⁰Including sets containing multiple similar objects.

⁵¹This may in fact be never, if the field has an `onDemand` hint and is not used.

- `const` and `auto` fields
- Reading, Writing and Appending of files compatible with the specification.
- Deferred Deserialization
- Restrictions: Singleton, Default, Nullable
- Hints: Lazy, Owner, ReadOnly, Ignore
- Full
 - Interfaces To do (154)
 - Enumerations To do (155)
 - other Restrictions
 - other Hints
 - Proper treatment of shadowed field definitions as of section 5.7.
 - Support for files exceeding the guaranteed numerical limit range⁵².
 - Safe behavior of interleaving states.
 - Language dependent treatment of comments, e.g. integration into doxygen or javadoc.⁵³
 - Name mangling to allow for usage of language keywords or illegal characters (unicode) in specification files, without making a language binding impossible.
 - A public reflective interface allowing reading of instances and type information of unknown types. To do (156)

F Numerical Limits

In order to keep serialized data platform independent, one has to respect the numerical limits of the various target platforms. For instance, the Java Virtual Machine can not deal with arrays with a size larger then about 2^{31} . Therefore we establish the following rule:

(De-)serialization of a file with an array of more then 2^{30} elements or a type with more then 2^{30} instances may fail because of numerical limits of the target platform. Although, strings can have at most 2^{32} bytes of data, strings are usually represented as array of characters, therefore their length is expected to not exceed 2^{30} characters as well.

G Numerical Constants

This section will list the translation of type IDs (as required in section 10.4) and restriction IDs (see section 7.1 and 10.4). Restrictions with even IDs can be ignored if unknown. To do (157)
To do (158)
To do (159)

⁵²depending on the target platform, this might not be desirable or might even be a for-free feature

⁵³This may even require a language extension providing tags inside comments which are translated into tags of the respective documentation framework.

Type Name	Value	Hex Value		
const i8	0	0x0		
const i16	1	0x1		
const i32	2	0x2		
const i64	3	0x3		
const v64	4	0x4		
annotation	5	0x5		
bool	6	0x6		
i8	7	0x7		
i16	8	0x8		
i32	9	0x9		
i64	10	0xA		
v64	11	0xB		
f32	12	0xC		
f64	13	0xD		
string	14	0xE		
T[i]	15	0xF		
T[]	17	0x11		
list<T>	18	0x12		
set<T>	19	0x13		
map<T ₁ , ..., T _n >	20	0x14		
T	32 + idx_T	0x20 + idx_T		

(a) Type IDs

Restriction Name	Value
nullable	0
range	1
unique	2
constantLengthPointer	3
singleton	4
coding	5
monotone	6

(b) Restriction IDs

Glossary

base type The root of a type tree, i.e. the farthest type reachable over the super type relation. 35

built-in type Any predefined type, that is not a compound type, i.e. annotations, booleans, integers, floats and strings. 17, 18

subtype If a user type A extends a type B, A is called the sub type (of B). 13, 36, 46, 50, 51

supertype If a user type A extends a type B, B is called the super type (of A). 13, 32, 36, 37, 46

unknown type We will call a type *unknown*, if there is no visible declaration of the type. Such types must not occur in a declaration file, but they can be encountered in the serialization or deserialization process. 13

user type Any type, that is defined by the user using a type declaration. 13, 16, 18

visible declaration We will call a type declaration *visible*, if it is defined in the local file, or in any file transitively reachable over include directives. 53

Acronyms

- API** Application Programming Interface. 15, 29, 45, 48
- BPSI** Base Pool Start Index. 32, 36, 43
- EXI** Efficient XML Interchange Format. 6
- LBPSI** Local Base Pool Start Index. 32, 37
- SKiLL** Serialization Killer Language. 4–8, 12, 13, 15, 18–20, 24, 32, 34, 41, 48, 51
- v64** Variable length 64-bit signed integer. 17, 29, 31, 32, 36, 49
- XML** Extensible Markup Language. 4, 6
- XSD** XML Schema Definition Language. 6

References

- [Apa13] Apache Software Foundation. *Thrift Types*. <http://thrift.apache.org/docs/types/>, 2013.
- [Goo13] Google. *Protocol Buffers Language Guide*. <https://developers.google.com/protocol-buffers/docs/proto>, 2013.
- [GSMT⁺08] Shudi Gao, C. Michael Sperberg-McQueen, Henry S. Thompson, Noah Mendelsohn, David Beech, and Murray Maloney. W3c xml schema definition language (xsd) 1.1 part 1: Structures. World Wide Web Consortium, Working Draft WD-xmlschema11-1-20080620, June 2008.
- [iee08] IEEE Standard for Floating-Point Arithmetic. Technical report, Microprocessor Standards Committee of the IEEE Computer Society, 3 Park Avenue, New York, NY 10016-5997, USA, August 2008.
- [ISO11] ISO. *ISO/IEC 9899:2011 Information technology — Programming languages — C*. International Organization for Standardization, Geneva, Switzerland, December 2011.
- [jav13] *Javadoc Technology*. <http://docs.oracle.com/javase/7/docs/technotes/guides/javadoc/index.html>, 2013.
- [LA13] Chris Lattner and Vikram Adve. *LLVM Language Reference Manual*. <http://llvm.cs.uiuc.edu/docs/LangRef.html>, 2013.
- [Lam87] David Alex Lamb. Idl: Sharing intermediate representations. *ACM Trans. Program. Lang. Syst.*, 9(3):297–318, 1987.
- [LYBB13] T. Lindholm, F. Yellin, G. Bracha, and A. Buckley. *The Java Virtual Machine Specification: Java Se, 7 Ed*. Always learning. Prentice Hall PTR, 2013.

- [PGM⁺08] David Peterson, Shudi Gao, Ashok Malhotra, C. Michael Sperberg-McQueen, and Henry S. Thompson. W3c xml schema definition language (xsd) 1.1 part 2: Datatypes. World Wide Web Consortium, Working Draft WD-xmlschema11-2-20080620, June 2008.
- [RP94] J. Reynolds and J. Postel. Assigned Numbers. RFC 1700 (Historic), October 1994. Obsoleted by RFC 3232.
- [SK11] John Schneider and Takuki Kamiya. Efficient xml interchange (exi) format 1.0. Technical report, W3C - World Wide Web Consortium, <http://www.w3.org/TR/2011/REC-exi-20110310/>, March 2011.
- [TDB⁺06] S. Tucker Taft, Robert A. Duff, Randall Brukardt, Erhard Plödereder, and Pascal Leroy. *Ada 2005 Reference Manual. Language and Standard Libraries - International Standard ISO/IEC 8652/1995 (E) with Technical Corrigendum 1 and Amendment 1*, volume 4348 of *Lecture Notes in Computer Science*. Springer, 2006.
- [vH13] Dimitri van Heesch. *Doxygen User Manual*. <http://www.stack.nl/~dimitri/doxygen/manual/>, 2013.
- [xml06] Extensible markup language (xml) 1.1 (second edition). W3c recommendation, W3C - World Wide Web Consortium, <http://www.w3.org/TR/2006/REC-xml11-20060816/>, September 2006.

To do...

- ☐ 1 (p. 1): create a pygments skill grammar and remove listings
- ☐ 2 (p. 1): nur falls das kommt(sieht nicht danach aus): Fritjof Alexander Mayer zu additional critics für die i64 in felddaten
- ☐ 3 (p. 1): Die Spezifikationssprache sollte Grundtypen, wie bisher verwenden; in der serialisierung sollte es aber nicht so sein; eventuell will man später sowas wie "Matrix f32[4][4] data;" erlauben.
- ☐ 4 (p. 1): Irgendwo sollte man unterbringen, dass der usability gedanke dazu führt, dass die spezifikationssprache und das binärformat eher losgelöst voneinander existieren
- ☐ 5 (p. 5): howto read this document: user, full reference, implementer
- ☐ 6 (p. 5): über die vier teile von skill reden: spec, api, internal state, sf $\rightarrow \exists$ foto
- ☐ 7 (p. 5): sicherstellen, dass lazy und deferred jetzt immer on demand genannt wird
- ☐ 8 (p. 6): Hier was über API/Spec/SF/Binding sagen.
- ☐ 9 (p. 6): Jedes Kapitel vorstellen und erklären was es beinhaltet und wen es vermutlich interessiert.
- ☐ 10 (p. 7): cite GXL, besonders die absurd verschwenderische und unlesbare repräsentation
- ☐ 11 (p. 7): cite TGraphs (nicht nur xml); das scheint wirklich anwender zu haben: Graph Technology in Reverse Engineering The TGraph Approach Jürgen Ebert Volker Riediger Andreas Winter; hier könnte ein skill backend zeigen, was man hier gewinnen kann
- ☐ 12 (p. 7): note: TGraphs benutzen scheinbar ein eigenes textbasiertes format, das nicht xml-basiert ist.
- ☐ 13 (p. 7): cite XMI, ähnlich absurd
- ☐ 14 (p. 8): Hier sollte man noch erwähnen, dass sich die "architecture of computer systems" in den letzten 25 Jahren so sehr verändert hat, dass annahmen über die effizienz konkreter strukturen nicht mehr ohne weiteres richtig sind; hier kommen vor allem memory mapped files rein und eine erwähnung der puffer und blöcke hierarchie auf der man sich mittlerweile sogar beim ram bewegt

- ☐ 15 (p. 8): some people seem to realize that the in-memory representation can be moved over the network without modification, if they do not have to deal with endianness and stuff like that
- ☐ 16 (p. 8): CapNProto zitieren <http://kentonv.github.io/capnproto/index.html>
- ☐ 17 (p. 8): SBE zitieren <http://weareadaptive.com/blog/2013/12/10/sbe-1/>
- ☐ 18 (p. 8): citation needed?
- ☐ 19 (p. 8): cite,cite,cite
- ☐ 20 (p. 8): cite,cite,cite
- ☐ 21 (p. 8): section on notation; hier sollte man auch erwähnen, dass sf namen in beispielen zu binardateien in der testsuite gehören
- ☐ 22 (p. 8): falls sinnvoll zu jedem beispiel eine referenzdatei erzeugen
- ☐ 23 (p. 9): ensure correctness of last sentence!
- ☐ 24 (p. 9): (ensure and state that) all examples in this publication are legal skill words; do this by creating test for every one of them
- ☐ 25 (p. 10): maybe map should be map<groundtype, maptype|groundtype>
- ☐ 26 (p. 10): das frontend soll eine option zur verfügung stellen, die see annotationen in kommentare einfügt, die auf die spezifikationsdatei verweisen, damit man die spezifikation bei größeren systemen wieder findet
- ☐ 27 (p. 14): semantik von namensräumen aufschreiben
- ☐ 28 (p. 16): comment tags shall be treated case insensitive and may be followed by an optional ":" that is to be removed from the text
- ☐ 29 (p. 16): [...] außerdem soll nach zeilenumbrüchen ein gemeinsamer präfix verwendbar sein um kommentare alignen zu können. Newlines müssen sich im parser wieder finden; aus kompatibilitätsgründen muss es aber auch eine plain version geben, die keine newlines enthält und alles in einem string zusammenfasst
- ☐ 30 (p. 16): die letzten sätze ersetzen; hier muss bedeutung soll wie in ... sein stehen
- ☐ 31 (p. 16): comment token: leading stars and whitespace, as well as trailing white space will be removed. If two words are separated by whitespace, they shall be separated after trimming of whitespace as well.

- 32 (p. 16): "the comment system was designed with doxygen and javadoc in mind; ref both
- 33 (p. 17): auto fields müssen bei @mixin nicht vorhanden sein, da hier kein code generiert wird und der wert von auto feldern für die serialisierung und den state vollkommen gleichgültig ist; hier muss man darauf achten, dass auto-erreichbare instanzen nicht über closures in den state gelangen, sonst passieren merkwürdige dinge
- 34 (p. 17): Note: There can at most be one file per type containing an unmodified or "==" modified type declaration of that name. Note: "++" and "--" must not contradict each other. Note(style!): There **should** at most be one file containing change modifiers for a given name space. Note: Change modifiers are processed in the front-end. Note: "==" modified declarations have to be compatible to an unmodified appearance of the same type definition.
- 35 (p. 19): compatibilitätsregel: kleinere integer werden immer akzeptiert, größere werden nur dann akzeptiert, wenn ihre werte in den erwarteten bereich passen
- 36 (p. 20): cast regel: falls eine datei eine annotation enthält aber ein typ erwartet wird, so wird die datei akzeptiert, falls alle instanzen dem erwarteten typ genügen; in diesem fall ist ein append unzulässig
- 37 (p. 20): hier fehlt ein beispiel und eine begründung wofür annotations gut sind
- 38 (p. 20): durch die dazugekommene cast regel kann man annotations in der migrationsphase nutzen um sich mit manchen typen erst auseinanderzusetzen, wenn sie tatsächlich gebraucht werden
- 39 (p. 20): hier muss man noch mal darauf hinweisen, dass sich formate in einem ständigen fluss befinden, mit dem man umgehen muss
- 40 (p. 20): annotations dürfen beliebige typen ersetzen, solange diese kompatibel sind; das führt bei containern o.ä. dazu, dass das feld niemals einen nicht-default wert enthalten darf, weil sonst der cast fehlschlägt. ⇒ annotations sind quasi das top element des typhalbverbands
- 41 (p. 21): this is no longer the case; strings must be unified if they reside in memory at the same time, but this leaves potential copies

of user strings; this is a bit more relaxed than unique, but it is necessary, because otherwise all strings had to be deserialized eventually.

- ☐ 42 (p. 21): idea only(dont do it, just imagine): remove the ground type concept; make maps binary
- ☐ 43 (p. 21): the problem with having containers as base types is that they would not have a reference semantics but a value semantics, which might be very confusing especially for java programmers; the old way is the better solution; we should however keep the n-ary maps a concept of the specification language and force the binary files and the api to treat them as right associative binary maps
- ☐ 44 (p. 21): note: having compound type arguments is a bit like free variables in logic, because users might think that $T[]$ for any T is a Type, like T itself, but that's not the way skill treats types
- ☐ 45 (p. 21): make clear that containers are stored in a flat way into an object. This is surprisingly what most users require and there HAS to be a way to do this in the current encoding schema. If one wanted to share such objects, they can be moved into their own type declaration.
- ☐ 46 (p. 21): $\text{map } a \ b \ c = \text{map } a \ (\text{map } b \ c)$
- ☐ 47 (p. 21): interfaces; interfaces can either be used to group properties of subtypes or to describe properties of types; both can not be mixed, because the mapping onto regular types would no longer be monotone (the intersection of two unrelated supertypes is an annotation, but super types would be "deeper"; that's a problem)
- ☐ 48 (p. 21): interfaces as subtypes of regular types are a promise, that all implementations are subtypes of the regular type as well; therefore only one regular type can be inherited
- ☐ 49 (p. 22): nach der erweiterung müssten hier deutlich mehr regeln stehen
- ☐ 50 (p. 22): type order: interfaces sind kleiner als klassen (damit v64 kleiner sind)
- ☐ 51 (p. 22): ad interfaces: in der IR geplättete interfaces sind ein usability Problem, weil man aller Wahrscheinlichkeit nach ungeprüfte typecasts im binding als user braucht, weil man keinen typ hat, der die existenz eines feldes garantiert; das sollte man eher vermeiden

- ☐ 52 (p. 22): interfaces dürfen nur kommentare haben, aber weder hints noch restrictions
- ☐ 53 (p. 22): ad enums: wenn man defaults hat sind enums nurnoch darstellungssache; im sf sind sie so teuer wie normale integer, weil sie als benannte subtypen per reference quasi über ihren ordinalen wert serialisert werden, was der C lösung entspricht, nur dass man die usability von Java enums bekommt:)
- ☐ 54 (p. 22): move and distribute
- ☐ 55 (p. 23): @diss: hier schreiben, dass IML voll solcher entscheidungen ist
- ☐ 56 (p. 23): move and distribute
- ☐ 57 (p. 24): erklären, was genau typsicher für uns bedeutet; eventuell auch etwas über zustandsübergänge sagen, weil das auch etwas anders ist als in üblichen sprachdefinitionen
- ☐ 58 (p. 24): es gibt kein write! es gibt nur state management funktionen, wie compress, die auf einer leeren datei das selbe sind, wie früher write
- ☐ 59 (p. 24): type order: \Rightarrow es wäre nett, wenn man beweisen kann, dass man type order in $O(n)$ bucket-sorten kann
- ☐ 60 (p. 24): erklären, dass container nur 64bit größe haben, d.h. sie können nur entsprechend viele instanzen beinhalten und index-typen sollten uint64 sein
- ☐ 61 (p. 24): there shall be a both, a pragma restriction and a pragma hint; both intended for exploratory purpose of the skill format; a warning shall be emitted upon use of the front-end if a pragma is encountered to prohibit use in production code
- ☐ 62 (p. 25): das muss alles etwas strukturierter werden; außerdem sollte man subparagraphs für die 4 kategorien einfügen und erklären, was sie bedeuten; insbesondere sollte man bei der serialisierung erwähnen, dass es ein vorgriff auf das nächste kapitel ist
- ☐ 63 (p. 25): skippable restrictions may not have a serialization other than epsilon
- ☐ 64 (p. 25): unknown restrictions obtained from reading a file have to be treated as if they were specified; the only exception to this rule is the remove hint

- ☐ 65 (p. 25): es sollte erwähnt werden, dass empfohlen wird restrictions erst beim serialisieren oder explizit zu prüfen, da es sonst umständlich bis unmöglich werden kann typen zu instantiieren; hier sollte man ein beispiel mit nonnull graphen machen
- ☐ 66 (p. 25): hier muss man noch was zu fehlerbehandlung sagen, falls sich restrictions widersprechen, was vor allem bei default und range passieren kann; das sollte aber in der praxis bedeuten, dass die anwendung mit den daten sowieso nicht arbeiten kann
- ☐ 67 (p. 25): add an unsupportedRestrictionException, if a restriction can not be parsed; the message shall contain at least the involved type/field, location and index. There shall be an option for the generator, which causes the generated code to raise this exception if the file would be parseable as well
- ☐ 68 (p. 27): wie bei feldern, nur dass es für alle instanzen gilt; eine etwaig vorhandene feld restriction überschreibt den default des typs
- ☐ 69 (p. 27): man sollte erwähnen, dass field restrictions signaturen haben, je nach argument typ
- ☐ 70 (p. 28): translation required!
- ☐ 71 (p. 28): (hier muss man die skill-spec anpassen)
- ☐ 72 (p. 28): hier sollte eine liste von default werten für den fall stehen, dass keine default restriction angegeben ist; die antwort ist im prinzip immer 0, es sei denn es gibt kein 0
- ☐ 73 (p. 28): consider: The section about default values somehow disappeared from the TR. Default values need to be specified again.

To do (??)

In contrast to the old definition, which was basically equivalent to the java byte code default values definition, compound types shall have empty collections as default instances. This has no impact on the ABI, because serialization of empty collections and null pointers is equivalent.

- ☐ 75 (p. 29): doch, bitvector!
- ☐ 76 (p. 30): hier (oder vermutlich in einem anhang) sollte man als Beispiel "bitvector" verwenden, um sets<T> oder bool[] zu kodieren
- ☐ 77 (p. 33): Das hier ist in wirklichkeit ein coding "i64"

- 78 (p. 33): mixins erklären; das sollte den platform mechanismus abbilden; man muss aber platformspezifische escapes definieren; hier sollte man vielleicht eine *.language.mixin format als seiten-spezifikation definieren um ausreichend customization zu bieten; anm.: mixins sind ein sehr fortgeschrittenes feature, sowohl für binding, aber vor allem für den user
- 79 (p. 34): auch deserialization? eigentlich sollte unifikation durch den hint vorgenommen werden und von der restriction erhalten
- 80 (p. 36): no: its a list of arrays-> selber memory footprint, aber (konstant) mehr indirektionen
- 81 (p. 36): state somewhere, that although skill is casing agnostic, casing will influence the generated binding
- 82 (p. 36): this chapter requires a rewrite
- 83 (p. 36): section state transitions: create, read, write, append; hier sollte man einen graphen malen, welche übergänge zulässig sind; außerdem sollte man erwähnen, dass weder write noch append daten wegwerfen dürfen
- 84 (p. 36): reorganize this whole section; it should explain exactly how skill states should work, how state transitions are realized and how implementation can be done in parallel(on SMP nodes)
- 85 (p. 36): falls das noch nicht so ist, dann muss man hier fordern, dass die datenstrukturen in einer inlinefreundlichen weise organisiert werden, aber nicht direkt exportiert, damit man bei änderungen, wie z.B. !distributed nicht den anwendungscode ändern muss, sondern die arbeit vom compiler für einen übernommen wird
- 86 (p. 36): es gibt im prinzip zwei ansätze: den kernel/typecontext (generischer) und den generated; beide haben vor und nachteile; der kernel ist nur möglich, wenn es inlining zur linktime gibt, sonst macht er performance bedingt keinen sinn
- 87 (p. 36): SKill IDs dürfen nur für serialisierte daten vergeben werden; das sorgt dafür, dass der umgang einheitlich bleibt
- 88 (p. 36): Pools sollten auf ihren state verweisen, damit man auch mit einem einzelnen pool in der lage ist den kompletten state zu speichern; scheint auch nutzbarer zu sein
- 89 (p. 36): states sollen drei funktionen zur verfügung stellen: check, closure und flush; check testet restrictions, closure fügt alle transitiv erreichbaren instanzen dem state hinzu und flush macht check

und closure, löscht alle als gelöscht markierten instanzen und schreibt die änderungen auf die platte

- ☐ 90 (p. 36): es muss eine state.add und eine state.delete methode geben; jede instanz sollte zudem eine delete methode haben und jeder typ eine add methode, die aber an denn state weiterleitet (andernfalls ist der typ nicht notwendigerweise korrekt abgebildet)
- ☐ 91 (p. 36): die kernel sache funktioniert nur, wenn man auch in der lage ist die speicherrepräsentation zu kontrollieren und damit auch zu inlinen; das könnte in java trotzdem funktionieren, wenn die deoptimization zuverlässig genug arbeitet oder trotz jar nur eine anbindung existiert; dann wäre es als lösung für den JIT overhead aber grob gelogen
- ☐ 92 (p. 36): das lese/schreibe API sollte sich an C-Style files orientieren; man sollte den mode beim öffnen wählen, es sollte außerdem eine flush und eine close methode geben; außerdem gibts hier ein paar weitere, wie z.B. ein explizites check
- ☐ 93 (p. 36): ein State sollte sich wie ein File und wie eine Collection aus Typen verhalten
- ☐ 94 (p. 36): Typen sollten sich wie eine Collection aus Instanzen und eine Collection aus Feldern verhalten
- ☐ 95 (p. 36): Felder verhalten sich wie eine Collection aus den Feldwerten der entsprechenden Instanzen
- ☐ 96 (p. 36): hier sollte man etwas über die "hairy type hierarchy" sagen: es hat jeder bekannte typ einen untyp, der alle unbekannten untypen repräsentiert
- ☐ 97 (p. 36): hier sollte man was über 64bit unsigned integer indices sagen
- ☐ 98 (p. 39): in wirklichkeit sind storage pools natürlich keine (resizable) arrays, sondern resizeble arrays aus festen arrays, damit man sich beim resizen die kopie spart; irgendwas in der richtung sollte man machen. es ist nämlich nicht möglich, dass man ein objekt verschieben muss; außerdem sollte man löschen sowieso erst zum Zeitpunkt der serialisierung festklopfen, oder wenn der user flush() sagt
- ☐ 99 (p. 39): find a better quote and move this one to the motivation text

- ☐ 100 (p. 39): every example should have a binary skill file in the test suite and the name should be included in the example
- ☐ 101 (p. 39): strings als typnamen sollten bis auf das eine feld im typeblock verbannt werden; dadurch kann man die uniqueness von strings loswerden und muss nicht mehr immer alle strings deserialisieren, was in der praxis ein größeres problem sein sollte
- ☐ 102 (p. 39): cite bauhaus
- ☐ 103 (p. 39): als allererstes sollte direkt vor diesem kapitel erklärt werden, was ein abstrakter state und was abstrakte objekte sind, damit man hier eine notation hat um serialisierung zu erklären; da sollte auch sowas drin stehen wie σ ist die menge aller U und u die menge aller felder
- ☐ 104 (p. 40): talk about state transition system with transitions: c,r,m,a,w,order
- ☐ 105 (p. 40): die regel mit aufnehmen, dass felder, die nur default werte haben nicht serialisiert werden, wenn sie zuvor nicht existierten. das ist wichtig um z.b. in IML tote datenformatsteile los zu werden
- ☐ 106 (p. 40): strings dürfen zusammengelegt werden, müssen aber nicht; das verhindert, dass man beim schreiben strings aus der datei lesen muss
- ☐ 107 (p. 40): enums brauchen in der serialisierung präfixe, sonst kollidieren sie eventuell mit anderen typen; \Rightarrow enum namen wie namespaces behandeln!
- ☐ 108 (p. 40): serialisierung von refs wird auch konsequent über indices gemacht!
- ☐ 109 (p. 40): IDs werden von 1 gezählt; 0 bei super bedeutet kein supertype, 0 bei index ist reserviert um zur laufzeit die skillID zu repräsentieren
- ☐ 110 (p. 40): fields have to be written in ascending order
- ☐ 111 (p. 40): das wort (L)BPSI sollte aus der arbeit verbannt werden; man sollte von offsets reden und die offsets sollten so gewählt sein, dass er für den basepool implizit 0 ist und keiner verschiebung bedarf; man sollte ein bild malen, dass die verschiebung der subPools zeigt

- 112 (p. 40): typblock kann so umsortiert werden, dass man keine preliminaryTypes mit indices mehr braucht: \Rightarrow typblock in type-order sortieren \Rightarrow felder in einen zweiten teil abtrennen, der genauso geordnet ist wie der typblock und direkt dahinter kommt
- 113 (p. 40): wenn man die offsets in einen weiteren teil versprengt, dann kann man beim schreiben einen bereich mappen; man müsste mal schauen, wie sich das auswirkt; vielleicht will man für die end offsets tatsächlich i64 verwenden, dann könnte man die auch parallel schreiben
- 114 (p. 41): field IDs müssen zu jedem zeitpunkt dicht sein und beginnen bei 0 aufsteigend
- 115 (p. 42): integrate nicely:
- 116 (p. 42): reintroduce the notion of type order and explain, why it simplifies (de-)serialization
- 117 (p. 42): In case of offsets, they shall be chosen in an ascending manner.
- 118 (p. 42): Does the scala binding conform the todo above and why is this the case? (may require reordering of pools array?)
- 119 (p. 43): ensure that the whole file does not contain the word "lazy"
- 120 (p. 43): ensure that this chapter is no plain contradiction to statements above
- 121 (p. 45): check against (closed) bug nr 13 on github
- 122 (p. 45): -felder mit nur null/default referenzen dürfen nicht serialisiert werden -pools ohne echte instanzen dürfen nicht serialisiert werden \Rightarrow die kombination aus beidem erlaubt letzteres (was in tr13 nicht geht)
- 123 (p. 46): ein beispiel machen, mit leerem set auf einen typ, der samt set deswegen nicht referenziert wird
- 124 (p. 46): prüfen! das sollte sowieso BPO lauten
- 125 (p. 47): write something about sub-storage pools
- 126 (p. 47): make sure this is consistent with the actual file format
- 127 (p. 48): hier ganz besonders aufpassen!

- ☐ 128 (p. 48): wenn die typen anders serialisiert werden, dann braucht man insbesondere bei maps einen test; das führt dazu, dass alte dateien mit neuen nicht mehr kompatibel sind!
- ☐ 129 (p. 48): bool muss man ändern, als true -> 0x01 bis 0xff, damit man bei der serialisierung konvertierung sparen kann; die meisten implementierungen deserialisieren sowieso schon mit != 0 oder > 0
- ☐ 130 (p. 49): umbenennen! z.b. "age"
- ☐ 131 (p. 49): call it age; use @min(-1) and document that by saying -1 is the age of an unborn child
- ☐ 132 (p. 49): phase 1: collect instances
- ☐ 133 (p. 49): phase 2 : instances are, already in pools, but age restriction has to be checked!
- ☐ 134 (p. 49): this can be heavily simplified by calling the current stream *s*.
- ☐ 135 (p. 50): this is no longer correct!!!
- ☐ 136 (p. 51): das ganze kapitel muss man neu machen; viele von den aussagen stimmen so nicht, weil die deserialisierung etwas komplizierter wird
- ☐ 137 (p. 51): das LL-Parsing Kapitel sollte man streichen; statt dessen sollte man ein Kapitel mit Implementierungstipps basteln, das gute Ideen aus veröffentlichungen sammelt und zusammenfasst.
- ☐ 138 (p. 53): hier könnte man eventuell flussdiagramme verwenden; das ist leichter zu verstehen und man braucht nicht den formalismus des states
- ☐ 139 (p. 54): das hier ist jetzt genauso age; wenn man vorne was ändert muss man das hier berücksichtigen, insbesondere wenn man eine restriction verwendet
- ☐ 140 (p. 55): lazy age example
- ☐ 141 (p. 55): mehr praktische nutzung muss abgewartet werden
- ☐ 142 (p. 58): remove specification of namespaces; note: keep selector in default, because enums create mini namespaces; enums verwenden den selben encoding mechanismus wie namespaces für instanznamen

- ☐ 143 (p. 58): copy grammar for fields and restrictions
- ☐ 144 (p. 60): note that encode is interpreted as unsigned, wherever used internally and as signed, wehere used explicitly by the user
- ☐ 145 (p. 60): url writer
- ☐ 146 (p. 60): url reader
- ☐ 147 (p. 60): diese datei sollte man auf teilen
- ☐ 148 (p. 60): check error reporting to reflect the changes in skills behavior
- ☐ 149 (p. 60): note that the earliest point to check restrictions is at the end of a type block; suggest that applicability of restrictions shall be checked after each block (to indicate the tool violating the restriction) and semantic checks shall be done as last action of the read operation in order to allow for as much parallelism as possible
- ☐ 150 (p. 60): constant have to be checked as soon as they are encountered
- ☐ 151 (p. 60): This is less restrictive then claiming that no offset may point after the EOF. This way partially lost data can still be processed.
- ☐ 152 (p. 61): das hätte schon länger field lauten müssen, oder?
- ☐ 153 (p. 61): cite students:)
- ☐ 154 (p. 62): einführen, in das alte loch der dependent arrays stopfen und bool-arrays konstanter länge verbieten; hier muss auch noch ein aufrunden auf i8[...] spezifiziert werden
- ☐ 155 (p. 62): einführen, über implizite supertype, iml-style als subtyp eines echten typs
- ☐ 156 (p. 62): einführen, über singletons und einen gemeinsamen subtype; hier sollte die spezifikationssprache einen mechanismus bieten enums wegzuprojezieren
- ☐ 157 (p. 62): rename: fully generic file manipulation?
- ☐ 158 (p. 62): should be two tables(type and field)
- ☐ 159 (p. 62): add additional serialization column
- ☐ 160 (p. ??): das ist nicht mehr richtig!