Report

An IDE-based, integrated solution to Schema Evolution of Object-Oriented Software

Author(s):
Piccioni, Marco; Oriol, Manuel; Meyer, Bertrand; Schneider, Teseo

Publication Date:
2009

Permanent Link:
https://doi.org/10.3929/ethz-a-006836319

Rights / License:
In Copyright - Non-Commercial Use Permitted
An IDE-based, integrated solution to Schema Evolution of Object-Oriented Software

Marco Piccioni∗, Manuel Oriol†, Bertrand Meyer∗ and Teseo Schneider∗
∗ ETH Zurich, Switzerland
Email: {marco.piccioni,bertrand.meyer}@inf.ethz.ch, teseos@student.ethz.ch
† University of York, United Kingdom
Email: manuel@cs.york.ac.uk

Abstract

With the wide support for serialization in object-oriented programming languages, persistent objects have become common place. Retrieving previously "persisted" objects from classes whose schema changed is however difficult. Naive solutions, such as initializing new fields to default values, risk to invalidate the consistency of the application. The ESCHER framework addresses this issues through an IDE-based approach that handles schema evolution by managing versions of the code and generating transformation functions automatically. The infrastructure also enforces class invariants to prevent the introduction of any corrupt objects. This article describes the principles behind invariant-safe schema evolution, and the design and implementation of the ESCHER system.

1. Introduction

The most widespread approach to handling object-oriented schema evolution relies heavily on class developers: they have to provide conversion code to import older object schema versions into the new ones. To make an informed decision, developers must have the previous versions of the class available and prepare conversion code for all versions of the class. Designing such code is not trivial. Many currently available persistence solutions are coupled with tolerant retrieval algorithms, automatically making questionable decisions about the initialization of the retrieved objects. This enforces silent acceptance of possibly inconsistent objects into the system. This is a recipe for disaster, as class invariants may be silently violated. Following Bloch [7], retrieving objects previously stored should ensure establishing the class invariant and that no illegal access to the object is possible.

To tackle these issues, we first study how serializable classes from the java.util package evolved from Java 1.2.2 to Java 6.0. It appears that 17.5% of the changes have impacted the persistence of the serializable classes of the package. This allows the identification of the most frequently occurring refactorings. We then propose a formal model that represents the refactorings as well as the corresponding transformation functions.

We propose to shift the focus of developers from runtime to development time by introducing ESCHER (Eiffel SCHema Evolution suppoRt), a modified EiffelStudio IDE. The tool aims at guiding and supporting develop-
ers through the schema evolution process at class release time, includes version handling and code template generating features for the transformation functions. In addition we provide a robust retrieval algorithm to prevent the acceptance of inconsistent objects into the system.

Section 2 presents the analysis of some newly collected data about the refactorings of java.util. Section 3 describes the model for software updates. Section 4 details the implementation of both the IDE integration and the retrieval algorithm. Section 5 analyzes contributions and limitations of the current approach. Section 6 summarizes the previous approaches, from both the authors and others. Finally, section 7 describes our conclusions and future work.

2. Schema Evolution in Practice

Advani et al. [5] already evaluated refactorings1 in fifteen open source Java systems. These studies show that refactorings like “rename field”, “move field”, “rename method”, and “move method”, account for approximately 66% of the total refactorings identified. We can also observe that the “rename field” and “move field” refactorings alone account for 32% of the total refactorings identified. While these results are calculated on all classes of the considered systems, there is no evidence that classes meant to be persistent would exhibit the same characteristics.

To check if persistent classes evolve in a similar manner, we consider the serializable classes from the Java package java.util. Analyzing Java code was a conscious choice because classes that might have persistent instances are Serializable. The package java.util was chosen because it contains classes whose instances are likely to be serialized. In fact the package contains classes that model collections, dates, currencies, and locales. We considered the 22 classes in the package directly implementing the Serializable interface and analyze them manually across five major versions of the language: 1.2.2, 1.3.1, 1.4.2, 5.0, 6.0. We also take into consideration 22 refactoring types, ten of which are directly relevant to the serialization process. These 22 refactorings are shown in Table 1 and were created by using a systematic approach that considered all possible changes (addition, removal, modification) over all possible targets (attribute, visibility, methods etc). Note that the complete raw data is available for download [6].

The list of refactorings we considered as relevant for Java persistence is the following:

- Attribute added
- Attribute removed
- Attribute renamed
- Attribute type changed
- Attribute initialization value changed
- Attribute becoming a constant
- Constant becoming an attribute

These refactorings are a superset of the usual ones [14] and aim at expressing more precisely the semantics of evolving classes whose objects are intended to be serialized. The “Attribute initialization value changed” refactoring is considered because it may influence the class invariant. The “Attribute becoming a constant” and its counterpart “Constant becoming an attribute” are relevant because constants are typically not serialized.

In line with what was discovered previously, Table 1 shows that the persistence-related refactorings consist of 17.5% of the total number of refactorings.

Table 2 shows how refactorings are distributed across different classes and across all five language versions analyzed. The data suggests that persistence-related refactorings are sufficiently widespread among the classes, and confirm that persisted data might actually

---

1. In this article, consistently to the work from Advani et al. [5] we use the term refactoring for any minor modifications of the code rather than semantics preserving modification-only.
change significantly over time.

Table 3 shows the number of refactorings per kind, considered across all versions. Note that “Attribute added” and “Attribute removed” together constitute 74% of all persistence-related refactorings.

What this short study shows is that classes whom instances might be serialized change over time. Moreover 17.5% of these changes directly impact the capability of classes to deserialize the instances of their previous versions. This is aggravated by the fact that there may be a high number of stored objects that need to be updated. This does not directly imply that deserializing will be performed in a semantically inconsistent way, but it is likely that it will create issues at some point.

In the next section, we show a model based on this analysis that allows to represent refactorings and to generate converters.
3. Model

This part presents a model of updates that particularly emphasizes the generation of conversion functions — contrary to other previous models of refactoring [1], [2].

The upper part of Figure 1 presents a simplified syntax of class definitions in Eiffel programs. We omit the declaration both of routines and constraints on generic parameters as they are not included in the serialized form. We also do not explicitly consider inheritance because we have access to the flattened version of the class. This is a valid assumption as serialized objects are de facto flattened as well.

The following subsections define refactorings and class transformations, how to generate a type transformer from a class transformation, and possible strategies to extract type transformers.

3.1. Classifying software evolution: refactoring units

We call refactoring a basic transformation applied to a class. A refactoring is a function modifying at most one attribute:

\[ R : \text{class} \mapsto \text{class} \]

The bottom part of Figure 1 shows the definition of the five kinds of refactoring.

Taking into account the data analysis in the previous section, we define five standard refactorings that our system recognizes:

- Attribute not changed
- Attribute added
- Attribute renamed
- Attribute type changed
- Attribute removed

They are sufficient to semantically include all the refactorings listed in the previous section. The three missing refactorings are treated as follows: “Attribute value changed” is taken into account when evaluating the new class invariant clause, so no special action is needed. “Attribute to constant” is considered as “Attribute removed”, as constants are not serialized. “Constant to attribute” is considered as “Attribute added” for the same reason as above.

A class transformation \( T_{R_1,...,R_N} \) going from one version of a class to another can then be described by a list of refactorings \( R_1,\ldots,R_N \) (\( N \geq 1 \)):

\[ T_{R_1,...,R_N} : \text{class} \mapsto \text{class} \]

such that:

\[ T_{R_1,...,R_N}(\text{class}_{0}) = (R_n \circ \ldots \circ R_1)(\text{class}_{0}) \]

Note that any modifications to the attributes of a class can be described by some class transformation — for example by deleting all attributes whose names are not present in the class anymore and adding the new ones. Thus, class transformations are complete with respect to attribute modifications.

While there is always a decomposition, using a straightforward algorithm might not produce the best results and devising heuristics is essential to use this model as the basis of an implementation.

3.2. Object transformers

While refactorings allow an easy representation of the static transformations of a class, this is not sufficient to generate the object transformers that create an instance of the new class from a serialized instance of the old class. The main reason is that there is a need for an explicit default initialization of new attributes. To define such initializations, the programmer actually needs to input default initialization values. To represent that, we use the token \( I[a:] \) (for input), evaluating into the next value in a list of inputs. We also add an output generator \( W[a:] \) (for warning), warning programmers of the impossibility of generating a translation or a removal operation and evaluates into \texttt{noop}. 
The upper part of Figure 2 presents a succinct syntax of object transformers. The generation of object transformers from a class transformation can be expressed as transformation functions as shown in the bottom part of Figure 2. The transformation function $G[\cdot]$ takes a class transformation and generates the associated object transformer. The current solution relies on statically comparing the abstract syntax trees (AST) of both classes to detect refactorings. This implies that in certain situations, like the “Attribute renamed” refactoring below, the outcome of the tool can only suggest the possibility that the specific refactoring has happened. The extraction relies on a set of heuristics successively applied:

1) An attribute that does not change name and declared type generates an “Attribute not changed” refactoring. The code generator assigns a release number to every class and creates a system release. No other action is necessary.

3.3. Heuristics for schema evolution refactorings

While we are currently considering the extraction of transformations based on actual changes made by programmers using the IDE,
2) An attribute that does not change name but changes type between two versions generates an “Attribute type changed” refactoring. Here the code generator provides the best results, being always able to both detect the refactoring and provide a complete code generation.

3) An attribute $att$ in the old version that does not have a counterpart with the same name in the new version while having at least a counterpart $att'$ with the same type in the new version is a candidate for generating an “Attribute renamed” refactoring. By only comparing the two classes’ ASTs it is not possible to determine what happened in this specific case. The code generator, after detecting this refactoring, will generate a conversion function template warning about the possibility of a rename. A better way of analyzing the code to find out about that point would be to check whether the variables are used by the same clients in the same context. As it requires a global analysis of the code we considered it was too time consuming and could actually be captured in a better way by tracking user direct use of a renaming function.

4) An attribute of the new version for which a counterpart with the same name cannot be found in the old version generates an “Attribute added” refactoring. The framework always detects this refactoring, but the only reasonable suggestion that can be offered is to initialize the attribute with the default value and to give the hint to check the new class invariant with respect to the old one. Here we do not see space for a real improvement, because the initial value of a newly added attribute
is largely determined by a semantics known to developers only.

5) An attribute of the old version that cannot find a counterpart in the new version generates an “Attribute removed” refactoring. The code generator always detects this refactoring, and issues a warning when in the new version there is a new attribute having the same type as the removed attribute. In fact this could be a possible case of renaming.

To detect the refactorings we iterate through the new class attributes, search for a correspondence with attributes in the old class and create a corresponding heuristic. We then repeat the process starting from the old class, in order to gather more information, for example to find all the attributes that were removed.

Here is an example that can help clarifying the code generation step. Figure 3 shows the code of a simplified version of a class BANK_ACCOUNT. Note that the balance is not stored as such but computed on-demand. Imagine that for efficiency reasons we decide to change the implementation of the query balance to use a field instead of computation. A new attribute balance is then created and is updated every time a deposit or a withdraw takes place. Clients are not affected in this case, because in Eiffel both attributes and functions can be accessed in the same way from outside the class, providing uniform access to the class itself. In addition, the attribute info is stored in a more appropriate numeric field. The resulting class is illustrated in Figure 4: The conversion code will be generated in an ad-hoc generated class, (see Figure 5), containing all the conversion functions necessary for converting objects of different versions into one another.

The next section describes how we implemented this model and how it is integrated in the EiffelStudio IDE.

Figure 3. BANK_ACCOUNT, version 1.

4. Implementation

The ESCHER tool is integrated into EiffelStudio, but can be used also independently from it. It is flexible with respect to storage, because it allows developers to choose between the “physical” and a “logical” representation of an object. The “physical” representation encodes every object detail appearing in the corresponding class structure. This is clearly more sensitive to the smallest modifications of the structure itself. However, there may be information that is not useful, safe, or efficient to store. As suggested by Bloch [7], the ideal serialized form of an object contains only the logical data represented by the object. Finally, it keeps the
class BANK_ACCOUNT
inherit VERSIONED_CLASS
create make
feature --- Version implementation
  version : INTEGER
  --- Class version number.
do
  Result := 2
end
feature --- Initialization
  make
  --- Creation feature.
do
  balance := 1
end
feature --- Status report
  balance : INTEGER
  --- Account balance.
  info : INTEGER
  --- Some numeric information.

invariant
  valid_account : balance > 0
end

Figure 4. BANK_ACCOUNT, version 2.

draft evolution code cohesive, because all the conversion code necessary for migrating instances between different versions of the given class is located in only one ad hoc handler class, different from the class itself. Though our approach can be applied to any object-oriented programming language providing support for storing and retrieving objects, we use Eiffel mainly because of its integrated support for Design by Contract, and in particular for class invariants. Class invariants occupy a very important role in complementing the work done by the code generator [4]. Figure 6 shows a screen capture of the ESCHER tool with the following elements: (1) the buttons to trigger a new release, create and release a schema evolution handler class and create a filter class to customize serialization; (2) the different release folders and the schema evolution handlers folder on the right; (3) the serializer and serializer code generator libraries on the right.

The versioning mechanism relies on the notion of release. A system release is a versioned set of classes compiled and thus released together. While each different class has a different class version identifier, different versions of the same class can only be part of different releases. At release time, ESCHER automatically increases the class version numbers of the modified classes. At runtime, objects of two different versions of the same class will never coexist. Provided the right conversion functions are there it will always be possible to retrieve any object of a certain class and version into an object of another version of the same class. This applies to both forwards and backwards updates as customers might be running an old system (specified by an old release number) and need to retrieve objects stored by a newer system release.

While modest, the support for generating code provided by ESCHER relieves developers from writing boiler plate code to transform instances stored in another version to the current one. At runtime, ESCHER relies on an ad hoc written library which automatically
performs the conversions across two different versions. It is important to note that the algorithm will raise an exception when one of the following conditions is not met: (1) the specific schema evolution handler exists, (2) the specific conversion function between the two versions exists, or (3) every specific field converter exists.

Figure 7 presents a simplified description of the algorithm expressed in pseudo-code. The algorithm is part of a serialization/deserialization library which is decoupled from the IDE, and therefore separately usable. The detailed documentation about the mechanism, a step-by-step tutorial, the source code and the executables for the tool are available for download on the ESCHER project page.²

if (old_ver_obj.is_equal (new_ver_obj))
perform standard_retrieval
else if (not class_version_handler_exists)
raise exception ("Conversion impossible for this class")
else if (not version_handler_exists (v1, v2))
raise exception ("Incompatible versions")
else if (not attribute.is_convertible)
raise exception ("A specific attribute cannot be converted")
else perform cross_version_retrieval

Figure 7. Algorithm for updates.

5. Evaluation

This work proposes a shift of attitude on how developers presently cope with class schema evolution. More precisely, by integrating support into an IDE, the approach

elevates class schema evolution to the status of first-class citizen of the software development process rather than undesirable side effect of the software production activities. It also proposes a significant time shift for any schema evolution effort, moving it from runtime to release time. A system release becomes then an event triggering a whole set of tool activities intended to help developers focus more on possible issues that may arise from previous versions of the newly released code. By analyzing some data about a widely used Java library, we have shown that schema evolution happens also for persistent classes.

To evaluate ESCHER we check how ESCHER scores with respect to the 113 persistence-related refactorings previously detected. This can be done because the considered refactorings are clearly language-independent and they can be easily recreated using ESCHER. The result obtained is that ESCHER is able to exactly identify the correct refactorings in all cases (100%) from the changes in `java.util`. Even if this is unlikely in the general case, especially for less stable code, this result suggest that the simple model we devised is enough to model a realistic set of refactorings. In Section 3 we have analyzed how each refactoring is recognized, and what can be done to further improve the framework support.

Threats to validity. By analyzing data from persistent classes in `java.util` it is clear that classes tagged as serializable evolve over time. The connection with persistent classes is however not that obvious as one could argue that serialization could also be used simply to send data over the network or for storing temporarily instances. It is therefore a threat to the validity of our data analysis.

While we think that ESCHER addresses most practical challenges, it still needs a
broader validation than the limited testing that we were able to perform. In particular the ease of use of the interface and the actual refactorings detection and code generation should be further investigated through extensive acceptance testing and the actual development and maintenance of applications that use persistence extensively. The ideal case would be that independent developers who use persistence in Eiffel would pick ESCHER to handle it. Not having such a study or such acceptance threatens the validity of our approach.

6. Previous Approaches

The most widespread approach relies on class modification to accommodate the changes in the class schema. The idea is to devise a class descriptor representing the information to serialize. To convert objects from an older version into the current ones, developers typically implement conversion functions. The Java language and the .NET framework with Version Tolerant Serialization (VTS) both follow this approach [10], [11]. The db4o OODB [12] advertises a “transparency” which seems overly optimistic, considering that retrieved objects with default values for attributes can be automatically accepted into the system.

The versioning approach, for example implemented in the Versant Fastobjects OODB, keeps the information about all the versions of each class. While providing a consistent view of the logical structure of the repository, it enables handling of backward and forward evolution of class schemata. Again, issues arise when dealing with a semantically consistent objects retrieval. To convert a stored object of a certain version into an object of the current version, CLOSQL [13] uses update or backdate routines. An unfortunate requirement is that a database administrator is needed every time a class is created, to specify which update or backdate routines have to be executed.

Orthogonal Persistence Java (OPJ) is an extension of the Java language Specification providing “orthogonal persistence” capabilities to the Java platform. The well known prototype implementation for OPJ, PJama, [8], [9] is an extension of the Java Virtual machine together with a persistent store, in which the state of an executing application is kept. The system state is checkpointed atomically and periodically to be able to recover from exceptions and crashes. PJama provides an approach to schema evolution that involves persisting both objects and classes, but again does not solve the fundamental consistency problem cited before. When a certain class evolves over time, it may be considered a different type, and named differently, or it may be considered the same type, leaving the same name but providing some other means of taking into account the different inner structure and semantics. While the second is mainstream, the first approach has been already explored [3] by the authors.

Another example of making different, parallel versions explicit can be found in UpgradeJ [16], though not focused neither on object persistence nor on enforcing semantical consistency. The automated detection of refactorings has been also extensively explored in literature, both with respect to existing libraries and software configuration systems, for example by Dig et al. [17].

The automatic generation of converters through type transformers generation has been described by Neamtiu et al. [15]. It focuses on updating structs in C programs whose layout might evolve. However, it is not per se linked to object-orientation and it does not benefit from having a model or an integration into an IDE.

3. www.versant.com (last visited: 15/5/09)
7. Conclusions and Future Work

Classes which produce persistent instances evolve over time. In our study more than one change out of six impacted the compatibility of persistent instances. This high ratio implies that the basic mechanisms that support persistence — like serialization— need a way to handle the schema evolution of their classes.

This article presents the ESCHER platform to cope with this issue. It relies on an IDE which identifies refactorings and generates semi-automatically migration code. The practical results recognizes all changes made in serializable classes from java.util over the past 5 major releases of Java and would generate conversion functions automatically.

Thanks to the use of class invariants and seamless integration into the development life cycle, ESCHER takes a step forward towards a more reliable schema evolution handling, and seems a valuable addition to the current scenario of software development.

In the near future we plan to run usability studies with practitioners from industry and academia to further validate our approach.

References


