JOURNAL OF OBJECT TECHNOLOGY

Online at www.jot.fm. Published by ETH Zurich, Chair of Software Engineering. ©JOT, 2002

Vol. 5, No. 5, Special Issue: ECOOP 2005 Workshop FTfJP, June 2006

Why we should not add `readonly` to Java (yet)

John Boyland University of Wisconsin–Milwaukee, USA

In this paper, I examine some of reasons that “read-only” style qualifiers have been proposed for Java, and also the principles behind the rules for these new qualifiers. I find that there is a mismatch between some of the motivating problems and the proposed solutions. In particular, most have an overly restrictive “transitivity” rule, and all encourage “observational exposure” as a way to prevent representation exposure. Thus I urge Java designers to proceed with caution when adopting a solution to these sets of problems.

1 PROPOSALS FOR “READ-ONLY” IN JAVA

The purpose in having qualifiers such as `readonly` on types in a programming language is so that programmers can enlist the compiler (and loader) in enforcing rules about the proper use of data. One part of the program may be willing to grant access to data to another part of the program only if it can be guaranteed that the other part does not mutate the data. Several proposals have been made to add enforceable “read-only” qualifiers in Java programs:

JAC [24] Kniesel and Theisen’s system “Java with Access Control”;

Universes [28] Müller and Poetzsch-Heffter’s system for alias and dependency control.

ModeJava [37] Skoglund and Wrigstad’s mode system for read-only references in Java.

Javari [6, 41] Birka and Ernst’s system for Java with “Reference Immutability” updated in a more recent paper by Tschantz and Ernst. Where it is necessary to distinguish the proposals, they will be referred to as Javari$_1$ and Javari$_2$ respectively.

In this section, we compare these proposals with emphasis on their broad similarities. We also compare with “const” in C++ [40] and with our earlier paper on capabilities [12].

A motivating example is used in Section 2 which demonstrates short-comings of the “read-only” concept, in particular “observational exposure,” which is further

Cite this document as follows: John Boyland: *Why we should not add `readonly` to Java (yet)*, in Journal of Object Technology, vol. 5, no. 5, Special Issue: ECOOP 2005 Workshop FTfJP, June 2006, pages 5–29,
http://www.jot.fm/issues/issues_2006_06/article1

```
class B {
  A f1;
  readonly A f2;
  mutable A f3;
  mutable readonly A f4;
  readonly A method1() readonly {
    this.f1 = new A(); // Error: Writes field of readonly this
    this.f3 = null;
    if (...) return this.f1;
    else return this.f2;
  }
  A method2(A p1, readonly A p2) {
    this.f1 = p2; // Error: Field needs a read-write ref.
    this.f2 = p1;
    this.f3 = p1;
    return this.f4; // Error: Result needs a read-write ref.
  }
}
```

Figure 1: An artificial class illustrating read-only rules.

explained and criticized in Section 3. Section 4 describes some preferred alternatives to existing “read-only” proposals and other related work is reviewed in Section 5.

Basic Rules

In these systems, the type system is expanded to permit a reference’s type to be specified as `readonly`. Formally, if T is some class, then `readonly` T represents a super-type of the reference type T. In other words, it is permitted to implicit coerce a reference of type T to `readonly` T (widening conversion), but the reverse coercion requires an explicit cast. This rule carries over in the obvious way to parameters, return values, local variables and fields. The receiver of a method may be typed as `readonly`, in which we call the method a *read-only* method.

This change only applies to *references* to objects, not to the objects themselves. A single object may have both normal and readonly references to it at the same time.

The basic intuition is that a reference of “read-only” type cannot be used to change a field. For a quick example, consider the artificial class in Fig. 1. Ignoring the `mutable` annotation for now, this example illustrates the rule. In `method1`, the receiver is `readonly` (indicated directly before the body) and thus `this` has type `readonly B`. Thus it is not legal to write field `f1` here. However, we are free to *read* these fields and since the return type is a “read-only” type, we can return the contents of either field `f1` or `f2` without error.

The second method `method2` is declared normally (without `readonly`) and thus doesn’t have this restriction, but must still follow the types of the fields, and thus the write to `f1` fails because a “read-only” reference is an inappropriate value to store in the field; an explicit cast would be needed. However, the write to `f2` succeeds because the “read-write” parameter `p1` can be implicitly coerced into a “read-only” reference to be stored in `f2`. The write of `f3` does not require any coercion.

This example shows the difference between `readonly` which applies to the reference stored in a field (protecting the fields of the object referred to), and Java’s existing `final` annotation which refers to the field itself, making it non-updatable in any method. This is the distinction that C++ makes between a pointer to a “const” object (whose data members cannot be written) and a const data member with a pointer in it (the pointer can be used for mutation). C++ can make do with one keyword (albeit somewhat confusingly) because pointers are explicitly typed.

Java’s `final` annotation protects a field from being updated regardless of whether the method is read-only. JAC and Javari$_1$ include a `mutable` annotation (borrowed from C++) that has the opposite effect: it makes a field updatable, again regardless of whether the method is read-only. The same capability is available in Javari$_2$ using the keyword `assignable`. Thus we see that `method1` is permitted to update field `f3` despite being “read-only.” The `mutable` can be combined with `readonly`, and thus we see that `method2` cannot return the contents of field `f4` since a “read-write” reference is needed. Indeed Javari$_2$’s `assignable` (can be assigned by any method) is indeed the direct opposite of `final` (cannot be assigned by any method). The same would be true of `mutable` in JAC and Javari$_1$, were it not for transitivity, as explained in the next section.

Transitivity

In C++, some data members have pointer type, and others have object type. In a const object, an object data member is also const. In other words, if I have a pointer to a “const” object and get a reference to the object stored in this data member, this reference is also “const” (or rather it is also to a “const” object). This transitivity of “constness” is entirely reasonable since the second object is stored within the first. On the other hand, if the first object has a *pointer* to a third object, then if we fetch this pointer, there are no restriction on mutating this object. Again, this *lack* of transitivity is reasonable if we consider the state of an object to consist solely of what is stored inside it.

However, even in C++, some object’s state notionally extends to other objects stored on the heap. For instance a “map” object includes pointers to nodes in a red-black tree. Changing any of these nodes conceptually changes the map. And indeed a “map” will protect its nodes and “voluntarily” propagate “constness” to maintain desired invariants.

In Java, there is no possibility of storing one object inside another (at the language level), and thus all subsidiary objects must be referred to through (implicit) pointers. The question is then whether transitivity should apply or not. The *right* answer should depend on whether the referred to object is part of first object or not. If it is conceptually part of that object's representation, then transitivity applies, otherwise it does not. This is the rule used in "flexible alias protection" [29].

Unfortunately, of the four proposals described above, only Universes distinguishes the representation fields from other fields.[1] All four proposals decide that the safe approach is to assume all read-write fields in an object refer to its representation. This heuristic is reasonable for many classes, but notably not for container classes, since the elements in the container are not necessarily considered part of the container. For this reason, in our low-level capability system, "read-only" was specifically not transitive [12].

In the systems with `mutable` (JAC and both versions of Javari), this keyword cancels this transitivity. In JAC and Javari$_1$, this dual purpose for a single keyword is confusing, which is why Javari$_2$ reserves `mutable` only for cancelling read-only transitivity. One may wish that Javari$_2$ had left `mutable` with the C++-inspired semantics and invented a new keyword (such as `readwrite`) to cancel transitivity, but at least the separation of function is welcome.

In any of the four systems, transitivity is illustrated in that `method1` would not be permitted to fetch a read-write reference from field `f1`. The `readonly` annotation on the receiver carries over to the value fetched from a field of `this`. How then can we write a "getter" function for `f1` that returns a read-only reference if the receiver is read-only, but returns a read-write reference if the receiver is read-write? In Universes such a method is illegal since it exposes the representation of the object. JAC allows it with the rule that a "non-read-only" return type of a read-only method is understood as being linked with the actual receiver mode. In ModeJava, the annotation `context` on the return type indicates the same situation. In C++, the solution is to overload the method; the programmer writes two methods with identical (and short) bodies but with the two possible signatures. Javari (both early and recent) permits this solution, but also supports genericity using an extension to Java 1.5's generics (Javari$_1$) or a new keyword `romaybe` (Javari$_2$).

Dynamic casts

In C++, the programmer can "cast away constness" with impunity, which is in accordance with the general principles of the language favoring flexibility over safety.

In Java, where safety is much more important, a cast on a reference is used to perform a "narrowing conversion." At run-time, the object (if not null) is tested to see if it indeed is of the desired class, and if not an exception is thrown. So-called

[1] And Universes severely restricts the non-representation read-write references that an object may possess.

dynamic casts require that run-time type information be saved. In Java and in C++ objects with virtual methods, this comes at little cost. In the case of `readonly` the cost is non-negligible since it would involve keeping an extra bit in each pointer (similar to what is done in capabilities [12]). The potential cost (of a safe solution) is cited as the reason why JAC does not permit "read-only" to be cast away.

ModeJava supports something analogous to a dynamic cast to test whether the receiver of a read-only method is actually "read-write." Thus a read-only method may test its receiver and perform different actions depending on whether the receiver was actually "read-only" or not. This potentially surprising behavior weakens the meaning of the `read` method annotation and necessitates forced read-only conversions by clients that want a read-only semantics.

In Javari, casts have a peculiar if practical semantics. Dynamically casting a read-only reference as `mutable` always succeeds, but the pointer is marked as "read-only" so that if it is subsequently used for mutation, an exception is thrown. An analogy can made with the (statically unsafe) type rule that permits arrays to be implicitly coerced into an array of a super-type. If the array is used to store something inappropriate, an exception is raised. In both cases, a dynamic check on uses protects against the dangers of a permitted type-unsafe coercion. This rule is practical since it permits code using `readonly` to safely co-exist with legacy code that lacks proper annotations.

Unfortunately, this rule goes against the spirit of a dynamic cast (which is supposed to check the reference) and also means code may raise unexpected exceptions far from where the erroneous cast occurs. Java's `ArrayStoreException` is seen as a blemish on the language, required because of an over-permissive type rule. It would be unfortunate if adding "read-only" qualifiers required another such check. Furthermore implementing the cast in this way requires that (some) pointers include an extra bit.

One advantage of Javari's rule is that `readonly` is iron-clad: if a method takes a `readonly` receiver or parameter, then it absolutely cannot modify the state through the reference, even if the actual receiver or parameter is "read-write." This property compares favorably with `readonly` in Universes, as seen next.

Universes have yet another semantics for cast. The system distinguishes representation objects from other objects similar to ownership type systems. Every object has an owner. A read-only reference can be cast to "read-write" by its owner. Thus at run-time, the dynamic cast compares the owner of the object to the receiver of the current method, and succeeds if they are the same. The cost for handling checked dynamic casts is thus borne per-object rather than per-reference. A surprising implication is that a `readonly` reference is not really read-only, it simply cannot be used *externally* for mutation.

Summary

The four (or five) systems reviewed share most of the same rules in which `readonly` adds a layer on the type system. All assume (or require) that read-write references in a class are part of the representation and thus enforce transitivity of "read-only." Javari$_2$ has the most flexible semantics because it distinguishes what it calls "`mutable`" from what it calls "`assignable`." Dynamic casts (with varying semantics) can be supported at some additional run-time cost in object or pointer representation.

2 EXAMPLE

The desire for a "read-only" qualifier is motivated by considerations of software constructions. This section uses illustrative examples in Java where the lack of "read-only" exposes a module to misuse unless defensive programming is used. The examples chiefly come from a recent "read-only" proposal (Javari$_1$). In several of the cases, I argue that the underlying issue does not fit the idea of a read-only pointer.

The case for "read-only"

Figure 2 illustrates a number of situations where the programmer may wish the compiler would enforce good usage, when it does not.

1. The "intersect" method in its comment promises not to change (mutate) its parameter. However this commitment is not expressible in the signature of the method, and thus the compiler cannot be called upon to enforce the signature. In Javari (or using any of the other proposals, with perhaps minor notational changes), the signature could be written

   ```
   void intersect(readonly IntSet set)
   ```

   The compiler would then enforce that the parameter was not in fact mutated.

2. The next declaration of interest is the constructor. It accepts an array of integers and uses this array as its representation. This method exposes the representation to the client, since the client can retain the pointer to the array and then, for instance, set all the elements to zero, thus breaking the representation invariant. The client may even be unaware of this problem, assuming that the constructor would make a copy of the array. Birka and Ernst explain that if the array parameter were annotated `readonly` the compiler would notice the (now illegal) initialization of the field with the parameter and flag the error.

```
/** This class represents a set of integers. **/
public class IntSet {
  /** Integers in the set with no duplications. **/
  private int[] ints;

  /** Removes all elements from this that
   * are not in set, without modifying set. **/
  public void intersect(IntSet set) {
    ...
  }

  /** Makes an IntSet initialized from an int[].
   * Throws BadArgumentException if there are
   * duplicate elements in the argument ints. **/
  public IntSet(int[] ints) {
    if (hasDuplicates(ints))
      throw new BadArgumentException();
    this.ints = ints;
  }

  /** Number of distinct elements of this. **/
  public int size() {
    return ints.length;
  }

  public int[] toArray() {
    return ints;
  }
}
```

Figure 2: A partial implementation of a set of integers.
(Figure 1 from Birka and Ernst's "A Practical Type System and Language for Reference Immutability" [6])

```
public class IntSetView extends JPanel {
  private final IntSet model;
  /** Construct a view of the given integer set.
   * When the set changes, the client should
   * call repaint(). **/
  public IntSetView(IntSet set) {
    ...
  }
  ⋮
}
```

Figure 3: A class to view integer sets.

3. The caller of the `size` method is not expecting the method to perform a side-effect. In particular it should be possible to call this method even when one has a read-only integer set instance. For this reason, Birka and Ernst suggest using a `readonly` qualifier for the (implicit) receiver.

4. Turning to the `toArray` method, we see another case of representation exposure. The signature as given does not prevent the client from breaking the invariant by changing the values in the array.

   As Birka & Ernst observe, if the return value were designated `readonly` (and the compiler enforced this designation), the client would be unable to modify the array elements, and thus could not upset any invariants. The `toArray` method in the Java collection framework is supposed to return a separate (mutable) array. The use of `readonly` here would make it clear that this set does not conform to the framework, a useful result.

5. Finally, consider another class that maintains a graphical view of an integer set as seen in Figure 3. (The example here does not come from an earlier paper.) The class is not intended to modify the set, although it is expected to *view* modifications performed elsewhere. Using a `readonly` annotation on the set will ensure that the view behaves as expected in this regard.

In summary, a “read-only” qualifier enforced by the compiler can aid in preventing dangerous exposure and enable informal guarantees in comments about non-mutation to be made formal and checkable.

Shortcomings of “read-only”

I now go through the same examples again, and discuss some ways in which “read-only” captures only some of the intended properties.

1. Regarding the `intersect` method, there is an additional property of the method that most users would expect: that the method does not retain the reference to the parameter `set` after the method returns. Suppose the parameter were saved in order to "memoize" the intersection operation. There is a danger that changes in the set would invalidate the memo cache. Retention may also cause a space leak. A "read-only" annotation cannot prevent such retention.

2. The constructor takes an array. There are actually three different reasonable designs behind a constructor of this form:

   - The client is expected to release the array to the control of the ADT. In other words, the parameter represents the transfer of a "unique" pointer;
   - The array is intended to be immutable, not changed by either the ADT or the client.
   - The array is intended to be copied by the ADT, and not retained.

   None of these three situations is fully expressed by using a "read-only" annotation. In the first case, the array is intended to be mutable and thus cannot be protected by "readonly." In the second case, it is insufficient since it does not prevent the client from mutating the array. In the third case, the non-retention (as already explicated) is not expressed.

3. With the `size` method, the caller is again likely to assume that the reference to the receiver will not be retained.

4. With the `toArray`, using `readonly` to qualify the result prevents representation exposure (exposing the ADT representation to mutation by external agents), but the result would still permit *observational exposure* in which the ADT representation is visible to the outside, for reads only. As long as the set doesn't change, there is little problem, but when it does, the way in which the array is used will be visible. If the array were to be recycled and used in a different set, the client would notice surprising changes.

   In C++, an equivalent situation occurs with regard to iterators into vectors; one is not permitted to retain an iterator when the vector changes. This requirement cannot be expressed in the language (even though C++ has the "const" keyword). Java collections classes have a similar statically-unenforceable requirement.

On the other hand, in the case of the graphical view in Figure 3, we find that retention is indeed expected, even while the set is not assumed to be immutable. This example shows a case where the semantics of a "read-only" type qualifier fits the design intent well.

Thus sometimes, as seen in the final example, a "read-only" qualifier correctly expresses intent but frequently it does an insufficient job of expressing the intent and encouraging good software practice. The two issues of retention and observational exposure in particular are seen even in those examples used to motivate the addition of a “read-only” qualifier. This section has already pointed out some of the problems of retention. In the following section, I argue that observational exposure has a negative effect on software.

3 GLASS WALLS ARE NOT ENOUGH

Good software practice requires true privacy, not just lack of external change. The problem of *representation exposure* is well known, in which the (changing) internal representation of an abstract data type or object is made available to agents outside the implementation context. Less well known are the problems of *observational exposure* in which these outside agents are only permitted to read the data.

Representation exposure

Representation exposure is deleterious for modularity because it permits an external agent to mutate the representation state of an abstract data type or object. The implementation may require certain object invariants for correct functioning. Ensuring that these invariants are maintained is much more difficult when changes can occur from outside the implementation module. Section 2 gave an example of an integer set that uses an array of integers with the invariant that the integers in the array were all distinct. If the array is exposed, someone could easily break this invariant without being aware of it. As Leino argues [27], an invariant should only need to be checked in a scope in which it is known.

To the novice programmer, it may appear sufficient that representation fields in the object be declared private, but (in)visibility of names is not sufficient to prevent accessibility of data. *Aliasing*, in which the same piece of data can be accessed using different names may subvert the protection provided through naming. Thus in order to protect against representation exposure, it is necessary to check every piece of privileged code that assigns a pointer in the representation or accesses a pointer to the representation.

Aliasing is a very useful property and is intimately connected with the idea of *object identity*, that it matters which object one refers to, not just the object’s contents. Thus, I don’t wish to banish it from the language; rather it needs to be controlled. Now in some situations aliasing has entirely benign effects: when the state pointed to is immutable (cannot be changed and will not be changed) the various uses cannot conflict, they may even be unaware of each other, assuming automatic deallocation of memory.

One way to control aliasing of mutable state has been proposed through the use of ownership types (for instance the work of Clarke [17, 15] or Boyapati [10]). The *objects* (not just the fields) in the representation are indicated as such and are protected by a type system that does not permit access to representation objects from outside the owner. In these ownership type systems, all access are controlled; no distinction is made between reads and writes.

Other similar proposals (e.g. Universes [28]) note that an external agent cannot break an invariant with only read access and thus permit "readonly" references to representation state. Indeed, the relative safety of "readonly" pointers into the representation motivates even those "readonly" proposals that do not include an ownership-like system. However, even read access should not be granted, as I argue now.

Observational exposure

Observational exposure (in which read-only access is granted to *mutable* internal representation objects) has the following bad consequences:

- The inner workings of the "abstract" data type become part of its interface;
- The ADT may be observed in an "invalid" state;
- Concurrency errors may develop.

Each of these closely related points is expanded in the remainder of this section. I argue that one should hide the representation completely; "glass walls" that permit observation while protecting integrity are insufficient.

Observational exposure increases coupling.

If the inner workings of an abstract data type are seen by outsiders, they are no longer "inner workings" but rather part of the interface. If this observing is intended to do anything useful, then the states that are observed must be properly specified. Putting all this information in the interface will make it much more difficult to evolve the ADT, because changes will be resisted by clients.

The developer, on the other hand, may simply refuse to give the specification for the data observed. In this case, clients are likely to guess an API and make unwarranted assumptions. "That's not my problem," the developer of ADT may claim, but bad software structure is a function of the software system as a whole. Indeed observational exposure can be seen as the dual of classic representation exposure: in each case one part of the system is confused when properties of data they are observing change unexpectedly because of another agent mutating that data. In other words, the problem is due to a write access on one side of a supposed abstraction barrier interfering with read access on the other side.

Observational exposure may expose data in invalid states.

Another way of describing the problem of exposure is that either the entire representation is exposed, there is no abstraction barrier at all (in which case modification from the outside can be expected to maintain invariants), or else some aspects of the representation are still hidden. In the latter case, the observer of some of the representation sees an incomplete picture, which may appear to be invalid. Even if the part exposed is valid at the time of exposure, if the exposed reference is retained, the state it refers to may become invalid later.

For instance, suppose the `IntSet` kept a separate field of the number of elements in the array which are used. When the client asks for an array and this field is not equal to the length of the array, the array is first resized to fit the size of the set. So the client doesn't observe this extra behavior at first. But if (say) one element of the set is removed, the implementation may decide to continue using the same array, but not regard the last element part of the set. A client who has retained the array may see a duplicate element at the end of the array. The array will appear invalid.

Observational exposure may lead to concurrency errors.

If an object may be used (and changed) by multiple threads, it is important that the mutable state be protected by mutual exclusion locks. Lea describes the problem and solutions in detail [25]. Greenhouse describes the design needed and how Java code can be checked against a formally declared design [21, 19]. The recommendations are summarized here (Similar rules and checking are proposed by Boyapati and Rinard [9]):

Logically each piece of state is protected by a single lock object. Often that lock is the object whose fields are the state, but sometimes the lock is a completely different object. For instance an ADT may use a single (private) lock object to synchronize all accesses to fields of any of its internal objects. Or an ADT may use multiple locks to protect different groups of its objects.

Threads should synchronize on the protecting lock before accessing the data. This requirement applies not only to writes, but also to reads because otherwise one may observe the data in an invalid state. This invalid state may not only be due to transitory conditions during a mutation, but also because of memory system effects (because of local caching in a multiprocessor). For example, suppose that a method of the integer set ADT allocates a new array, initializes the elements and then assigns this array to the `ints` field. Another thread that does not properly synchronize *may* see the field change to point to a new array before the elements in that array are initialized.

Furthermore, if multiple locks are involved, and some actions on an ADT require accessing more than one, then it is essential that these locks be acquired in a fixed order or deadlock can easily ensue when threads access the locks in contrary orders.

If an ADT permits observational exposure, the client may not know what locks protect the state and what order in which to acquire locks. Not synchronizing on the lock or synchronizing on the wrong lock makes the results obtained almost meaningless, whereas synchronization in the wrong order may lead to deadlock. Thus observation exposure is only safe if the details of which objects protect which state is revealed and the required synchronization order is detailed. Such an interface may severly limit ADT implementation flexibility.

Summary

In conclusion, observational exposure, while less dangerous to an ADT than representation exposure, nevertheless yields improperly structured software. In fact, viewed from the vantage of the program as a whole, the problems are symmetric. Thus I submit we should not accept a proposal to solve the one problem while ignoring the other. Abstraction walls should be opaque.

In addition, most of the "read-only" proposals (with the exception of Universes, but including both Javari proposals) do not actually prevent representation exposure. They simply make it *possible* to limit representation exposure to observational exposure. Thus they should not be viewed as solving the representation exposure problem at all.

4 ALTERNATIVES

In this section, I describe some alternatives to address the software-engineering problems that various "read-only" proposals are addressing. The section starts with an alternative that simply tunes what I consider the current best proposal (Javari$_2$). Then it turns to established ownership systems. Finally, I describe how a new technique, "fractional permissions," solves the problems.

Improving Javari

The two problems with the various read-only proposals are

- They enshrine a transitivity rule rather than distinguishing representation fields from other fields. Furthermore, with the exception of Javari$_2$, there is no clean way to turn off the effect of that rule because of the multiple behaviors of the `mutable` keyword.
- They encourage the use of observational exposure as the way to solve representation exposure. In particular, "argument dependence" (unanticipated changes observed through a read-only pointer) can cause subtle problems that usually will not be noticed in unit tests, but occur only in system integration.

Tschantz and Ernst essentially concede the latter problem, but correctly point out that representation exposure is a more serious problem, and that one shouldn't require every language extension to solve all related problems [41]. Nonetheless, Javari (and most of the other proposals for "read-only") does not in fact solve the exposure problem; it merely provides a way to limit it to observational exposure.

However, "read-only" semantics are useful for program reasoning in any case, and thus the point that a language extension does not need to solve all problems is well taken, as long as the proposed extension does not make it harder to correctly address the other related problems. Here, argument dependence and transitivity are the main issues. Argument dependence is avoided by using method effect annotations (see later sections) rather than "read-only" annotations. Assuming `readonly` is added to Java, any system of annotations that intends to fix the argument dependence problem will need to reinterpret a "read-only" annotation as a method effect, which leads to a less natural semantics.

Transitivity is a problem if followed absolutely. Happily, Javari$_2$ avoids this problem by using two keywords: "`assignable`" to have the meaning of "`mutable`" in C++ (indicating fields that can be assigned from a read-only method) and "`mutable`" which means that a pointer is not subject to transitivity and is always "read-write." The object that such a field (in the latter case) might point to is presumably part of some larger abstraction. In ownership terms, its owner is (more) global. In such a case, the object should be read-only (its annotation notwithstanding) if its owner is in a read-only context. In this case, the better way to indicate the field is that its owner is unspecified, not that it specifically has a mutable reference.

Thus I recommend that if a proposal such as Javari$_2$ is accepted, the keyword "`mutable`" should be used for the purpose it has in C++ (indicating fields of no semantic significance) and not to mean that the owner is unspecified. Furthermore, an unannotated field should *not* be assumed transitive (that the objects it refers to are owned by this object), but rather than the owner is unspecified. Instead a keyword such as "`rep`" should be used for fields where transitivity applies. This will leave semantic space open for ownership which solves the exposure problem fully.

Furthermore, I would recommend that so-called "mutable" casts (which could normal Java syntax for casts) be treated in Java as a form of unchecked cast, leading to a warning as with Java 5's generics. The Javari alternative of run-time checking every field store seems untenable, and would lead to exceptions being thrown at locations far from the erroneous cast. This change is independent of the ones needed to avoid making later solutions harder.

Finally, one wonders why "`const`" is not the usual keyword suggested for "read-only" annotations. It has a long history, and reasonably well-defined semantics.

Thus in summary, a Javari-like "read-only" proposal would be best with the following changes:

1. Use "`mutable`" instead of "`assignable`."

2. Fields only follow transitivity if annotated “`rep`.”

3. Casting away “read-only” leads to a compile-time warning (only).

4. Use “`const`” instead of “`readonly`” as a keyword.

Ownership

An ownership type system tracks references and forbids aliases between logically separate components. Some well-known systems are those of Dave Clarke [15], Boyapati [8] and Aldrich’s ownership domains [2]. I omit mention here of Universes [28] since this proposal was discussed earlier and because it does not solve the observational exposure problem.

Traditionally there is a one-to-one mapping between an object and its ownership domain although Aldrich and Chambers [2] have generalized this. The fields of an object can point to “representation” objects (inside a domain). Ownership type systems don’t directly distinguish reads from writes, but ownership can be extended with effects systems, so that individual methods can indicate which domains are read or written [14, 39, 38].

Ownership type systems address both the transitivity problem and also the observational exposure problem:

1. Transitivity applies only to objects owned by a read-only object. In particular “representation” fields follow this rule. Moreover, with ownership types, even some return types can be inferred as read-only when the caller knows that the owner is readonly. Thus instead of a “mutable” loop-hole, we can correct indicate the context in which an object is to be considered read-only.

2. Ownership type systems disallow access to objects outside of their domains (or unless explicitly permitted [2]). Both read and write access are forbidden which means that observational exposure does not escape notice.

   If an object is intended to be accessed globally, it must be explicitly allocated in global domain, or else be immutable. A mutable object in a local domain simply cannot escape the domain.

Furthermore, ownership systems have been deployed in large software systems [1, 3] and thus have a proven track record.

Ownership systems in their simple forms forbid ownership transfer and all outside access to state, but research has shown that several additions work well with ownership. Borrowed pointers can be implemented using ownership polymorphism [5], and uniqueness can be added in a complementary way [3] or directly using ownership [16].

The steps to getting these abilities in Java include:

- Adding ownership. This step would solve the exposure problem immediately (with no need for "read-only" at all).
- Adding method effects to get the additional ability to limit modifications to state.

These steps have an additional advantage in that they produce annotations that can be given semantics using "fractional permissions" as described next.

Fractional Permissions

I argue that the real issue is access to mutable state. Access should be controlled. Thus I propose the use of "permissions" in which each piece of mutable state has a single concomitant "permission" that gives write access. This access can be "split" to give multiple read permissions, but write permission can only be restored when all read permissions are accounted for. Code is annotated with how the permissions are transferred and a static type system ensures that state is only accessed when in the possession of the requisite permissions. This idea is called "fractional permissions" [11].

For instance, a unique field is represented as a field that, as well as holding the pointer to an object, also holds all the permissions for accessing the state of that object. Thus no access to the object can be performed without using the field. Ownership is represented by storing the permissions to access the state of the object in its owner. Again, the object cannot be accessed except through its owner. Method effects are represented as permissions that are passed to a method allowing it to access the state that it needs to. When the method is done, the permissions are then returned to the caller.

If an enduring "read-only" reference is desired (as in the `IntSetView` example), one may use a read permission stored in a globally accessible location. When the viewer needs to update, it accesses that location (perhaps using synchronization) and retrieves the needed permission before accessing the set.

The permission type system is low-level, and thus I and others propose a system of annotations (uniqueness, effect, representation annotations) that is given semantics by being translated into permissions [13]. Since the permissions can be checked statically, there is no need to encumber the compiler back-end or run-time system with permission semantics.

The examples from Section 2 can be annotated with permission annotations described here. Figure 4 shows one reasonable possibility for annotating the code from Birka and Ernst. Figure 5 shows how one could annotate our own example code. Annotations are non-executable (do not affect run-time semantics).

- The `ints` field is annotated `@unique` which means that the field comes packaged with the permissions to use the array.

```
/** This class represents a set of integers. **/
public class IntSet {
  /** Integers in the set with no duplications. **/
  private @unique @in("All") int[] ints;

  /** Removes all elements from this that
   * are not in set, without modifying set. **/
  @reads("set.All") @writes("this.All")
  public void intersect(IntSet set)
  { ... }

  /** Makes an IntSet initialized from an int[].
   * Throws BadArgumentException if there are
   * duplicate elements in the argument ints. **/
  public IntSet(@unique int[] ints) {
    if (hasDuplicates(ints))
      throw new BadArgumentException();
    this.ints = ints;
  }

  /** Number of distinct elements of this. **/
  @reads("this.All")
  public int size() {
    return ints.length;
  }

  @reads("this.All")
  public @readsfrom("this.All") int[] toArray() {
    return ints;
  }
}
```

Figure 4: Example code from Fig. 2 annotated with permission annotations.

```
public class IntSetView extends JPanel {
  private final @in("All") @readonly IntSet model;
  /** Construct a view of the given integer set.  When the set
   * changes, the client should call repaint().
   **/
  @reads("set.All")
  public IntSetView(@readonly IntSet set) {
    ...
  }
  ⋮
}
```

Figure 5: Example code from Fig. 3 annotated with permission annotations.

- The `@in("All")` annotation means that this field (not the array stored in it) is nested within the (inherited) "All" *data group*, a grouping of state within an object [20, 26].

- The `intersect` method has an effect annotation that shows what state is read (the fields of formal parameter `set`) and written (the fields of the receiver). The previous `@in("All")` annotations enables the implementation to use the "this.All" permission to access the field `ints` and then to unpack the permissions for the array stored in this field so as to modify the array elements. Write permission always includes read permission because of our model of fractional permissions.

- The formal parameter `set` is *not* annotated and thus is passed without any permissions beyond what is available in the effects (which must be returned when the method returns). Thus we have effective non-retention.

- The constructor takes a "unique" array (to initialize its "unique" field). The formal parameter thus comes with its permissions. There is no representation exposure because the permission to access the array is not returned to the caller; the permission has been *transferred*. Constructors are implicitly also passed full permission for all the objects fields, and so this need not be declared.

- The `size` method has a read effect.

- The `toArray` method has a read effect and also indicates that the permission to read the resulting array is available only indirectly through that read effect. In other words, as long as the array is being used, the read permission passed to the method is inaccessible and thus no write permission can be formed. When the client is done with the array, its permission can be permitted to fall back into the read permission for the set, after which it cannot be retrieved.

Observational exposure does not occur because the ADT cannot mutate the array until its permission has been replaced.

The client must have the required permissions in order to call this method, and can reconstruct them when it is finished using the array.

- In Figure 5, the constructor takes a "readonly" set, one which is available globally. The permission to access this set is also needed to initialize the view. Henceforth, the view will use the global state to access the set.

Permissions can be used to model ownership domains as well as shared (globally accessible) state. And since it is possible to interpret existing ownership, uniqueness and effects annotations in a permission semantics, there is an easy upgrade path from a system using ownership and effects. This compatibility is helpful since implementing permission checking is still in the early stages [32]. We currently have a prototype and have only analyzed small programs. We are actively increasing our experience at this time.

5 OTHER RELATED WORK

Noble, Vitek and Potter in Flexible Alias Protection [29] noted the problem that a read-only reference could nonetheless cause dependency problems if the value could be mutated elsewhere and cause the read-only reference to return new values. They coined the term "argument dependence" to refer to this kind of coupling. These researchers also showed the importance of distinguishing the representation of an object from references to external objects unlike earlier alias control systems such as Islands [22] and Balloons [4]. The idea of designating the "representation" was developed further with Clarke [18, 17, 15]. Then with Drossopolou, Clarke showed how an ownership system could be used to encapsulate effects [14]. In these systems, the ownership system prevent any exposure of representation, and thus there was little need for "readonly." Clarke and Wrigstad [16] have shown one way how to integrate ownership and uniqueness, Aldrich and others [3] another.

Schärli and others [35, 36] define encapsulation policies so that dynamically typed languages can support different encapsulation policies for different users. Read-only abilities are defined by restricting access to methods that change state, a generalization of using special read-only interfaces. Unlike those earlier techniques, which rely on static typing, the policy is attached to the reference dynamically in the same way as a capability [12] includes both rights and the object pointer.

In an early paper [34], Reynolds defines "interference" as when a piece of mutable state is accessed by two supposedly separate parts of a program, with at least one part performing a write. O'Hearn and others revisited this work with a mixture on linear and non-linear logics (SCIR) [31]. The approach was changed from one which inferred effects to one in which statements were checked against allowable accesses. Write access was indicated in the linear part of the type, read access in the non-linear part.

Similarly, Walker and others [42] define a capability language (CL) in which linear (non-duplicable) capabilities indicate unique, mutable state and duplicable capabilities indicate shared, immutable state. As in our work, the permission is separate from the pointers that point to the state and alias types are used to connect the two. But using nonlinearity for read access does not permit a write capability in either SCIR or CL to be temporarily duplicated and then later returned to write status except in limited situations.

Fractional permissions [11] solved this problem by preserving linearity (read permissions are split, not copied). Bornat and others [7] have shown how the idea of fractions can be used with O'Hearn and Reynolds' separation logic [23, 30, 33] and indeed how the idea can be extended to handle "counting" as well as "splitting."

6 CONCLUSION

This paper reviews several proposals to add `read-only` to Java. I argue that the "transitivity" rule is too restrictive, and that "read-only" is insufficient to prevent deleterious observational exposure. Combined with an ownership system of some sort (or more generally with permissions) these problems could be overcome. Thus further consideration is required before adopting "read-only" qualifiers into Java.

Acknowledgments

I thank Bill Scherlis, Edwin Chan, Aaron Greenhouse, Tim Halloran and the others in the Fluid project at CMU for their collaboration and conversations. I thank Bill Retert, Scott Wisnieski, Tian Zhao and many anonymous referees for their comments on drafts. I thank Michael Ernst for providing advance copies of papers and for fruitful discussions at FTJP 2005.

REFERENCES

[1] Jonathan Aldrich, Craig Chambers, and David Notkin. ArchJava: Connecting software architecture to implementation. In *Proceedings of the IEEE International Conference on Software Engineering (ICSE '02)*, Orlando, Florida, USA, May 19–25, pages 187–197. ACM Press, New York, May 2002.

[2] Jonathan Aldrich and Criag Chambers. Ownership domains: Separating aliasing policy from mechanism. In Martin Odersky, editor, *ECOOP'04 — Object-Oriented Programming, 18th European Conference*, Oslo, Norway, June 14–18, volume 3086 of *Lecture Notes in Computer Science*, pages 1–25. Springer, Berlin, Heidelberg, New York, 2004.

[3] Jonathan Aldrich, Valentin Kostadinov, and Craig Chambers. Alias annotations for program understanding. In *OOPSLA'02 Conference Proceedings—Object-Oriented Programming Systems, Languages and Applications*, Seattle, Washington, USA, November 4–8, *ACM SIGPLAN Notices*, 37(11):311–330, November 2002.

[4] Paulo Sergio Almeida. Balloon types: Controlling sharing of state in data types. In Mehmet Akşit and Satoshi Matsuoka, editors, *ECOOP'97 — Object-Oriented Programming, 11th European Conference*, Jyväskylä, Finland, June 9–13, volume 1241 of *Lecture Notes in Computer Science*, pages 32–59. Springer, Berlin, Heidelberg, New York, 1997.

[5] David F. Bacon, Robert E. Strom, and Ashis Tarafdar. Guava: A dialect of Java without data races. In *OOPSLA'00 Conference Proceedings—Object-Oriented Programming Systems, Languages and Applications*, Minneapolis, Minnesota, USA, October 15–19, *ACM SIGPLAN Notices*, 35(10):382–400, October 2000.

[6] Adrian Birka and Michael Ernst. A practical type system and language for reference immutability. In *OOPSLA'04 Conference Proceedings—Object-Oriented Programming Systems, Languages and Applications*, Vancouver, British Columbia, Canada, October 26–28, *ACM SIGPLAN Notices*, 39(10):35–49, October 2004.

[7] Richard Bornat, Cristiano Calcagno, Peter O'Hearn, and Matthew Parkinson. Permission accounting in separation logic. In *Conference Record of POPL 2005: the 32nd ACM SIGACT-SIGPLAN Symposium on Principles of Programming Languages*, Long Beach, California, USA, January 12-14, pages 259–270. ACM Press, New York, 2005.

[8] Chandrasekhar Boyapati. *SafeJava: A Unified Type System for Safe Programming.* Ph.D., Massachusetts Institute of Technology, February 2004.

[9] Chandrasekhar Boyapati and Martin Rinard. A parameterized type system for race-free Java programs. In *OOPSLA'01 Conference Proceedings—Object-Oriented Programming Systems, Languages and Applications*, Tampa, Florida, USA, November 14–18, *ACM SIGPLAN Notices*, 36(11):56–69, November 2001.

[10] Chandrasekhar Boyapati, Alexandru Salcianu, William Beebee, and Martin Rinard. Ownership types for safe region-based memory management in Real-Time Java. In *Proceedings of the ACM SIGPLAN '03 Conference on Programming Language Design and Implementation*, San Diego, California, June 8–11, *ACM SIGPLAN Notices*, 38:324–337, May 2003.

[11] John Boyland. Checking interference with fractional permissions. In R. Cousot, editor, *Static Analysis: 10th International Symposium*, San Diego, California, USA, June 11-13, volume 2694 of *Lecture Notes in Computer Science*, pages 55–72. Springer, Berlin, Heidelberg, New York, 2003.

[12] John Boyland, James Noble, and William Retert. Capabilities for sharing: A generalization of uniqueness and read-only. In Jørgen Lindskov Knudsen, editor, *ECOOP'01 — Object-Oriented Programming, 15th European Conference*, Budapest, Hungary, June 18–22, volume 2072 of *Lecture Notes in Computer Science*, pages 2–27. Springer, Berlin, Heidelberg, New York, 2001.

[13] John Boyland and William Retert. Connecting effects and uniqueness with adoption. In *Conference Record of POPL 2005: the 32nd ACM SIGACT-SIGPLAN Symposium on Principles of Programming Languages*, Long Beach, California, USA, January 12-14, pages 283–295. ACM Press, New York, 2005.

[14] Dave Clarke and Sophia Drossopoulou. Ownership, encapsulation and the disjointness of type and effect. In *OOPSLA'02 Conference Proceedings—Object-Oriented Programming Systems, Languages and Applications*, Seattle, Washington, USA, November 4–8, *ACM SIGPLAN Notices*, 37(11):292–310, November 2002.

[15] David Clarke. *Object Ownership and Containment.* PhD thesis, University of New South Wales, Sydney, Australia, 2001.

[16] David Clarke and Tobias Wrigstad. External uniqueness. In Benjamin C. Pierce, editor, *Informal Proceedings of International Workshop on Foundations of Object-Oriented Languages 2003 (FOOL 10).* January 2003.

[17] David G. Clarke, James Noble, and John M. Potter. Simple ownership types for object containment. In Jørgen Lindskov Knudsen, editor, *ECOOP'01 — Object-Oriented Programming, 15th European Conference*, Budapest, Hungary, June 18–22, volume 2072 of *Lecture Notes in Computer Science*, pages 53–76. Springer, Berlin, Heidelberg, New York, 2001.

[18] David G. Clarke, John M. Potter, and James Noble. Ownership types for flexible alias protection. In *OOPSLA'98 Conference Proceedings—Object-Oriented Programming Systems, Languages and Applications*, Vancouver, Canada, October 18–22, *ACM SIGPLAN Notices*, 33(10):48–64, October 1998.

[19] Aaron Greenhouse. *A Programmer-Oriented Approach to Safe Concurrency.* PhD thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, 2003.

[20] Aaron Greenhouse and John Boyland. An object-oriented effects system. In Rachid Guerraoui, editor, *ECOOP'99 — Object-Oriented Programming, 13th European Conference*, Lisbon, Portugal, June 14–18, volume 1628 of *Lecture Notes in Computer Science*, pages 205–229. Springer, Berlin, Heidelberg, New York, 1999.

[21] Aaron Greenhouse and William L. Scherlis. Assuring and evolving concurrent programs: Annotations and policy. In *Proceedings of the IEEE International*

Conference on Software Engineering (ICSE '02), Orlando, Florida, USA, May 19–25, pages 453–463. ACM Press, New York, May 2002.

[22] John Hogg. Islands: Aliasing protection in object-oriented languages. In *OOPSLA'91 Conference Proceedings—Object-Oriented Programming Systems, Languages and Applications*, Phoenix, Arizona, USA, October 6–11, *ACM SIGPLAN Notices*, 26(11):271–285, November 1991.

[23] Samin S. Ishtiaq and Peter W. O'Hearn. BI as an assertion language for mutable data structures. In *Conference Record of the Twenty-eighth Annual ACM SIGACT/SIGPLAN Symposium on Principles of Programming Languages*, London, UK, January 17–19, pages 14–26. ACM Press, New York, 2001.

[24] Günter Kniesel and Dirk Theisen. JAC – access right based encapsulation for Java. *Software Practice and Experience*, 31(6), May 2001.

[25] Doug Lea. *Concurrent Programming in Java.* The Java Series. Addison-Wesley, Reading, Massachussetts, USA, second edition, 2000.

[26] K. Rustan M. Leino. Data groups: Specifying the modification of extended state. In *OOPSLA'98 Conference Proceedings—Object-Oriented Programming Systems, Languages and Applications*, Vancouver, Canada, October 18–22, *ACM SIGPLAN Notices*, 33(10):144–153, October 1998.

[27] K. Rustan M. Leino and Gren Nelson. Data abstraction and information hiding. *ACM Transactions on Programming Languages and Systems*, 24(5):491–553, September 2002.

[28] Peter Müller and Arnd Poetzsch-Heffter. A type system for controlling representation exposure in Java. In Sophia Drossopolou, Susan Eisenbach, Bart Jacobs, Gary T. Leavens, Peter Müller, and Arnd Poetzsch-Heffter, editors, *2nd ECOOP Workshop on Formal Techniques for Java Programs*, Nice, France, June 12. 2000.

[29] James Noble, Jan Vitek, and John Potter. Flexible alias protection. In Eric Jul, editor, *ECOOP'98 — Object-Oriented Programming, 12th European Conference*, Brussels, Belgium, July 20–24, volume 1445 of *Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg, New York, 1998.

[30] P. W. O'Hearn, H. Yang, and J. C. Reynolds. Separation and information hiding. In *Conference Record of POPL 2004: the 31st ACM SIGACT-SIGPLAN Symposium on Principles of Programming Languages*, Venice, Italy, January 14–16, pages 268–280. ACM Press, New York, 2004.

[31] Peter W. O'Hearn, Makoto Takeyama, A. John Power, and Robert D. Tennent. Syntactic control of interference revisited. In *MFPS XI, conference on Mathematical Foundations of Program Semantics*, volume 1. Elsevier, 1995.

[32] William Retert. Implementing permission analysis. Doctoral dissertation proposal, 2005.

[33] John Reynolds. Separation logic: a logic for shared mutable data structures. In *Logic in Computer Science*, pages 55–74. IEEE Computer Society, Los Alamitos, California, July 22–25 2002.

[34] John C. Reynolds. Syntactic control of interference. In *Conference Record of the Fifth ACM Symposium on Principles of Programming Languages*, Tucson, Arizona, USA, pages 39–46. ACM Press, New York, January 1978.

[35] Nathanael Schärli, Andrew P. Black, and Stéphane Ducasse. Object-oriented encapsulation for dynamically typed languages. In *OOPSLA'04 Conference Proceedings—Object-Oriented Programming Systems, Languages and Applications*, Vancouver, British Columbia, Canada, October 26–28, *ACM SIGPLAN Notices*, 39(10), October 2004.

[36] Nathanael Schärli, Stéphane Ducasse, Oscar Nierstrasz, and Roel Wuyts. Composable encapsulation policies. In Martin Odersky, editor, *ECOOP'04 — Object-Oriented Programming, 18th European Conference*, Oslo, Norway, June 14–18, volume 3086 of *Lecture Notes in Computer Science*, pages 26–50. Springer, Berlin, Heidelberg, New York, 2004.

[37] Mats Skoglund and Tobias Wrigstad. A mode system for readonly references. In *3rd ECOOP Workshop on Formal Techniques for Java Programs*, Budapest, Hungary, June 18. 2001.

[38] Matthew Smith. Toward an effects system for ownership domains. In *7th ECOOP Workshop on Formal Techniques for Java-like Programs*, Glasgow, UK, July 26. 2005.

[39] Matthew Smith and Sophia Drossopoulou. Cheaper reasoning with ownership types. In *Informal Proceedings of "International Workshop on Aliasing, Confinement and Ownership in object-oriented programming (IWACO)"*. Utrecht University, Netherlands, July 2003.

[40] Bjarne Stroustrup. *The C++ programming Language*. Addison-Wesley, Reading, Massachussetts, USA, third edition, 1997.

[41] Matthew S. Tschantz and Michael D. Ernst. Javari: Adding reference immutability to Java. In *OOPSLA'05 Conference Proceedings—Object-Oriented Programming Systems, Languages and Applications*, San Diego, California, USA, October 16–20, *ACM SIGPLAN Notices*, 40(10):211–230, October 2005.

[42] David Walker, Karl Crary, and Greg Morrisett. Typed memory management via static capabilities. *ACM Transactions on Programming Languages and Systems*, 22(4):701–771, 2000.

ABOUT THE AUTHORS

John Boyland is a professor in the Department of Electrical Engineering and Computer Science at University of Wisconsin-Milwaukee, USA. He can be reached at boyland@cs.uwm.edu.