Notes on the sources for New Foundations

Randall Holmes

October 30, 2024

I have elsewhere shown the consistency of the theory commonly called New Foundations or NF, originally proposed by W. v. O. Quine in his paper "New foundations for mathematical logic". In this note, I review that original paper and may eventually review some other sources one might consult for information about this theory[1]. Quine himself made some errors in this paper and later in his discussion of NF, and there are other characteristic difficulties that people have with this system which such a review might allow us to discuss.

Quine presents his original paper as an extension of the logicist program of Russell and Whitehead. His concern is to show that mathematical statements can be translated into statements of logic. It is important for him to note that it is sentences that are translated, and that it is not the case that a logical translation is provided for each symbol in a mathematical statement [though I would disagree with this emphasis: if the notions in NF are accepted as logical, then in fact every item in a mathematical statement admits a logical implementation, if one looks closely.]

I thought originally that the paper was light on means of inference but it actually presents an adequate set of axioms and rules. I will make at least one change in his definitions, update his notation in many places, and may suggest changes to his set of primitives, though they are adequate.

He makes it clear what he means by "logic", which is more that what is currently meant: he takes logic to include the theory of propositions, classes and relations in Russell and Whitehead. He explicitly notes that he is setting

[1]These might include Quine's book Mathematical Logic, Rosser's book Logic for Mathematicians, papers of Specker about ambiguity and the failure of AC in NF, and Jensens' paper on NFU.

aside the notion of propositional function (I am fine with that, though it is a very interesting notion).

He asserts that the primitives required are simply membership, alternative denial (the Sheffer stroke) and universal quantification. A caveat is that membership is not usually regarded as a logical notion, but this writer is sympathetic to the idea that it can be so regarded. Predication *is* a logical notion, and, especially in the context of New Foundations, membership can be regarded as an implementation of predication. He goes on to say that all of logic, and so all of mathematics, can be rephrased in terms of these notions. A supply of variables is required. If x and y are variables, $x \in y$ is a sentence of our language. If p and q are sentences of our language, $p|q$ is a sentence of our language. If ϕ is a sentence of our language, $(\forall x : \phi)$ is a sentence of our language. This is a complete description of how propositions in the primitive language of NF are constructed. Note that we use more modern notation for the universal quantifier, but this is an inessential change.

We insert some technicalities about free and bound variables (Quine does discuss this notion) and a precise definition of substitution in a modern style. An occurrence of a variable x in a sentence ϕ of our language as described above is *bound* if it appears as a component of a (not necessarily proper) component $(\forall x : \psi)$ of ϕ. Other occurrences are *free*. We write $\phi[y/x]$ for the result of replacing y with x in ϕ: this is a bit more complicated than simple typographical replacement of all occurrences of x with y. We define $z[y/x]$ as z if the variable z is distinct from the variable x and define $x[y/x]$ as y, We define $(u \in v)[y/x]$ as $u[y/x] \in v[y/x]$. We define $(p|q)[y/x]$ as $p[y/x]|q[y/x]$. We define $(\forall z : \phi)[y/x]$ as $(\forall z : \phi[y/x])$ if z is distinct from x or y and define $(\forall x : \phi)[y/x]$ as $(\forall x : \phi)$ and $(\forall y : \phi)[y/x]$ as $(\forall z : \phi(z/y][y/x])$, where z is a variable new to the entire context. More technicalities about substitution will be needed for more complex term constructions, but NF has no terms other than variables in its basic form.

Here we discover the first error in the paper, which had a major historical impact on this approach to set theory. He states that $x \in y$ is to be read "x is a member of y". He says, this makes sense only if y is a class, but goes on to say that if y is a non-class this can be read as $x = y$. He believes that he can harmlessly reduce true atoms (non-classes with no elements) to what are [because of this history] called "Quine atoms", classes which are their own sole elements. We remark that this is a mathematical error. To assume that the domain of non-sets can be implemented as Quine atoms is

in fact an extremely strong assumption in the context of New Foundations [more precisely, in the presence of stratified comprehension, which we will see defined below], and a very unlikely one: in particular, it allows the disproof of the Axiom of Choice, which is a very alarming outcome whatever one's feelings about this axiom. Quine had no way of knowing that Choice was impacted at this point in the development, but he really should have known that this was a mistake: we will discuss this at more length below.

He explains that $p|q$ neans "It is not the case that p and q are both true". This device (due to Sheffer) is a very economical way (in our opinion *too* economical) to provide a primitive covering all notions of propositional logic.

He explains that $(\forall x : P[x])$ means that $P[x]$ is true whatever x may be (for any value we may assign to x). This is a bit mysterious in a language where we have no names for these values we may assign, but Russell and Whitehead had already gone along this odd path.

Quine then proceeds to introduce familiar notations by definition. He defines $\neg p$ as $p|p$ (again, I am using more modern notation). He defines $p \wedge q$ as $\neg(p|q)$. He defines $p \to q$ as $p|\neg q$. He defines $p \vee q$ as $\neg p \to q$. He defines $p \leftrightarrow q$ as $(p|q)|(p \vee q)$. I have systematically changed his typography. This gives the usual vocabulary of propositional logic.

He defines $(\exists x : \phi)$ as $\neg(\exists x : \neg\phi)$.

He defines $x \subseteq y$ as $(\forall z : z \in x \to z \in y)$.

He defines $x = y$ as $(\forall z : x \in z \to y \in z)$. It is a significant difference from more modern treatments of set theory and indeed more modern treatments of NF itself that he takes equality to be a defined notion. This can be justified but it takes work. To get the standard properties of equality from this definition, one is placing a lot of confidence in having enough classes. We will discuss this at more length below.

He remarks at this point that unique eliminability of definitions is compromised by the need to supply a new bound variable here. He suggests providing a convention which exactly forces the choice of bound variable. One could also of course identify propositional notations which differ only by renaming of bound variables. These are technical issues which require real attention (they should not be shoved under the rug) but which are now well understood. We require explicitly that new bound variables introduced in definition expansions be typographically different from all other variables in the context. A suggestion is to require that any new variable that is introduced in an instance of a definition is the first new variable in alphabetical order which does not appear (free or bound) in its scope (as already ex-

panded in application of the definition). This can also be used to firm up the definition of substitution into quantified formulas. This is enough to make definitions give unique results, and it doesn't require more than lip service because renaming of bound variables is provably a valid move in our logic.

I should also remark that I am not following Quine in his admirable care in using metavariables distinct from variables. The original text can be consulted for that. The issues involved are again nothing to be shoved under the rug, but they are not part of my concern here.

The next notion which Quine introduces is improved if we make a correction, due (I believe) to Rosser. He introduces definite descriptions. We will write $(\theta x : p)$ for "the x such that p". Quine's language admits no terms. So (following Russell) we arrange for all uses of $(\theta x : p)$ to be eliminated by transformations depending on the context in which it is used. Quine finesses issues of scope by stipulating the context is always to be taken to be an atomic membership statement. We take $P[x]$ to be either $y \in x$ or $x \in z$, and expand $P[(\theta x : p)/x]$ (extending the definition of substitution into membership statements to complex terms in the obvious way, so this is either $y \in (\theta x.p)$ or $(\theta x.p) \in z$) as

$$((\exists u : (\forall x : p \leftrightarrow x = u) \wedge P[u/x])$$

$$\vee(\neg(\exists u : (\forall x : p \leftrightarrow x = u) \wedge (\forall v : (\forall w : w \in v) \rightarrow P[w/x])))$$

This differs from what Quine (or Russell) does in fine detail: we say that an atomic membership statement is true of $(\theta x : p)$ if it is true of the unique object such that p, if there is one, and otherwise is true of all universal classes [this theory proves that there is one and only one universal class]. The improvement is that this term always in fact refers to a unique object, so these terms have the same logical behavior as free variables, which is not true of Russell's descriptions. We use the universal class as our default object for the perhaps eccentric reason that this definition works in NFU without the need to introduce a specific constant $\emptyset$ for the empty set. One could use the empty set as the default object, of course, if one did have such a constant.

This change has no essential effect on the theory presented; it makes the means of inference when using terms easier to describe. It also makes it clear that adding definite description as a primitive construction would have no effect on the strength of the theory, while it might strengthen the logicist claim to implement all of mathematics, by supporting implementation of objects as well as statements.

Quine observes that the order of expansion of a membership statement between two definite descriptions might be taken to require attention, and indeed formally it does, but for our definition (and I believe not for his) the two expansions of $(\theta x : \phi) \in (\theta x : \psi)$ are always logically equivalent.

In all defined notions, descriptions can be put in any context which allows a free occurrence of a variable.

Quine then introduces $\{x : \phi\}$ as $(\theta A : (\forall x : x \in A \leftrightarrow \phi))$. The notation we use here is quite different from his but far more familiar. Note that $\{x : x \notin x\}$ is a perfectly good term, and it is provable on the basis of axioms for first order logic alone that it denotes (in effect) the universal class.

Then define $\{x\}$ as $\{z : z = x\}$ and $\{x, y\}$ as $\{z : z = z \vee z = y\}$. Note carefully that these definitions do not presume that there actually are such classes. No axiom of comprehension in general or pairing in particular has been introduced at this point.

Relations are to be defined as classes of ordered pairs, and the ordered pair (x, y) can be defined as $\{\{x\}, \{x, y\}\}$, though without evidence for existence of unordered pairs we do not really know that this is an ordered pair yet.

$\{(x, y) : \phi\}$ can then be defined as $\{z : (\exists x : (\exists y : z = (x, y) \wedge \phi))\}$: we can fluently describe classes representing relations. Our notation here is quite different from Quine's but will work. He remarks that relations with more than two arguments can be implemented as binary relations using ordered pairs, and does not discuss them further.

At this point he pauses and says that enough definitions have been given to provide access to the implementation of mathematics in logic given by Russell and Whitehead (we agree). Of course, we do not know yet that these definitions succeed, lacking comprehension axioms to firm things up.

Quine then presents formal rules for reasoning.

There is a single axiom, the axiom of extensionality:

$$x \subseteq y \to (y \subseteq z) \to (x = y)$$

There is a list of rules

R1: $((p|(q|r))|((s \to s)|((s|q) \to (p|s)))$ is a theorem for any propositions p, q, r, s

There was a typo in the 1937 paper, corrected in later printings of the paper.

R2: $(\forall x : \phi) \to \phi[y/x]$ is a theorem, where $\phi[y/x]$ is the result of replacing all free occurrences of x with y in ϕ [a defect in this formalization which Quine notes and corrects in a footnote is handled by our careful definition of substitution given above].

R3: if x does not occur in ϕ, and ϕ is stratified [this term is to be defined] $(\exists x : (\forall y : y \in x \leftrightarrow \phi))$

A $\in$-chain in a formula ϕ is a finite sequence of variables $\{x_i\}_{1 \leq i \leq n}$ such that $x_j \in x_{j+1}$ occurs in ϕ for each j such that $1 \leq j < n$. This is further called an $\in$-chain from x_1 to x_n of length n. A formula ϕ is *stratified* iff for any variables u and v occurring in p, all $\in$-chains from u to v are of the same length (if there are any).

A formula containing defined notions is stratified iff its definitional expansion is stratified.

R4: if p and $p|(q|r)$ are theorems, then q is a theorem.

R5: If $p \to \phi$ is a theorem, and x is not free in p, then $p \to (\forall x : \phi)$ is a theorem.

R1 is a sort of marvel: it is impressive to those who like Quine himself are overly impressed by clever minimal definitions.

It can be noted that with our contextual definition of $(\theta y : \phi)$, all instances of R2 with $(\theta y : \phi)$ in place of y are provable. This is not the case for the one Quine used. The logic of definite descriptions is transparent with this definition, and this allows sensible development of all sorts of term constructions. This works correctly for set abstracts in general, and for stratified set abstracts we can further note that the default case is never used.

Quine's rhetorical approach is different from ours: he starts out with unrestricted comprehension as R3, and then exhibits Russell's paradox and discusses ways to "fix" it. He discusses the restriction, which he advertises as a restriction of Russell's theory of types (a very late one, proposed only a few years before the New Foundations paper) which assigns natural number types to all variables and requires atomic formulas to be of the form $x^n \in y^{n+1}$ (other formulas being meaningless) and proposes R3 as we state it as a pragmatic simplication. We would prefer an approach which gives an inherent (in fact logical) motivation for the stratification criterion; we think that this is possible, though difficult, and would better support the actual primary purpose of the paper.

He discusses at length the "hall of mirrors" effect in simple type theory, the fact that all defined objects in the simple theory of types are endlessly reduplicated at higher types. He advertises as a virtue of NF the fact that all of these seemingly redundant copies become the same object.

I really like Quine's definition of stratification, which says nothing about assignment of types to any particular variable.

The definition of stratification can be extended to work freely with use of stratified set abstracts. This does not require any sort of complicated reasoning (as others have claimed: entire papers have been written about justifying set abstracts). Extend the definition of $\in$-chain in p so that one either has $x_j \in x_{j+1}$ occurring in p or x_{j+1} equal to a component $\{x_j : q\}$ of p. Then elimination of a set abstract has no essential effect on $\in$-chains at all: a set abstract is replaced by a fresh variable in any chains it participates in, with all conditions preserved, and the default case of definite descriptions causes addition of some unconnected and well-behaved new chains. Stratification of terms with set abstracts has a straightforward definition and is preserved when set abstracts are eliminated by contextual definition.

The last paragraph of the paper contains the second serious error. The existence of the empty set and its iterated images under singleton does *not* imply the axiom of infinity. The existence of the *set* containing these objects and no others would imply infinity and quite a lot more. This is an elementary logical error, and Quine surely at bottom knew better: what is shown here is that all models of NF are infinite, and that is much weaker than the assertion that it implies the axiom of infinity. NF *does* prove the axiom of infinity, but for different and rather alarming reasons. Jensen's NFU proves the existence of the empty set and each of its iterated images under the singleton operation, and is consistent with the assertion that the universal set is finite.

We return to the subtler initial error of assuming that strong extensionality can be postulated harmlessly. In a sense it is actually immediate and obvious that what Quine says is to be doubted. It is suggested to us that under certain conditions we replace $x \in y$ with $x = y$. But the relative types of x, y in $x \in y$ and $x = y$ are not the same. So this move is far from harmless.

Quine makes it clear in a footnote what he wants to do: he wants to collapse each atom together with its unit class (and so with all its iterated unit classes, a point which he does not explicitly make). This would be fairly easy to do in Zermelo style set theory, but basically impossible in New Foundations. If we begin with a theory with atoms, the idea is to redefine

$u \in v$ as $u' \in v''$ in the original sense of membership, where u' is u unless u is the iterated singleton of an atom a, in which case u' is a, and v'' is v unless v is the iterated singleton of an atom b, in which case $v'' = \{b\}$. This has exactly the desired effect: the iterated singletons of atoms are identified (they belong to the same sets as the atoms) and each atom becomes its own sole member. But it doesn't work, because to work as intended, the modified membership relation must have a stratified definition, so that it can itself be used in instance of comprehension, and it cannot: it depends essentially on recursion on the singleton operation. Quine apparently does not see the necessity of considering *iterated* unit classes in the treatment, or he is simplifying matters and thinking this will work by analogy with the situation in ordinary set theory.

There are other ways to effect the redefinition of membership which work in ordinary set theory, but all of them founder on the need for recursion on the singleton relation if one tries to implement them in NF.

Further, if this construction works, the original collection of atoms must be the same size as a set of singletons, and in fact the same size as a set of n-fold iterated singletons for each concrete n, and so in fact quite a small set. This is a very strong assumption. The models of NFU discovered by Jensen all have the collection of urelements far larger than the collection of sets, and these models also witness the fact that the modified definition of membership in the previous paragraph *cannot* have a stratified definition; it is not merely that we do not see one which might cleverly be defined later.

We review the definition of equality. What is needed from a definition of equality is that $x = x$ is a theorem (it is, with the given definition) and that if we have $x = y$ and $P[x/z]$, we can deduce $P[y/z]$ (the rule of substitution of equals for equals). There is a bogus argument for this: $P[x/z]$, so $x \in \{z : P\}$ so (by the definition of equality) $y \in \{z : P\}$, so $P[y/z]$. This is bogus because we do not know that $\{z : P\}$ has the intended extension unless P is stratified.

We can nonetheless verify the rule of substitution. We argue first that if P is a formula in which z occurs free exactly once, then $x \in \{z : P\} \leftrightarrow P[x/z]$. We prove this by induction on the structure of P, claiming and verifying at each step that there is a formal method to simplify $\{z : P\}$ to either the universal class or the empty set or the singleton of the universal class or a form $\{z : u \in v\}$, given information about the values of all parameters in P (actually, all we need is information about the truth values of finitely many sentences involving the parameters, so everything can be handled using reasoning about finitely many cases).

$\{z : a \in z\}$ and $\{z : z \in a\}$ have the intended extensions because these formulas are stratified.

$\{z : \phi|\psi\} = \{z : \psi|\phi\}$ so we can assume without loss of generality that the sole occurrence of z is in ϕ. If ψ is true (which we can determine from knowledge of the values of free variables in P other than z), $\{z : \phi|\psi\} = \{z : \phi\}$, which has the intended extension and simplifies as indicated by inductive hypothesis. If ψ is false, $\{z : \phi|\psi\}$ is the universal class. In both cases $\{z : \phi|\psi\}$ turns out to have the intended extension.[2]

$\{z : (\forall w : \phi)\}$ is naturally trickiest.

If $z = w$ this simplifies to the universe or the empty set depending on the truth value of $(\forall z : \phi)$.

If $\{z : \phi\}$ simplifies to the universal set or the empty class, or the singleton of the universal class, then $\{z : (\forall w : \phi)\}$ simplifies to the same thing in the case of the universal set or the empty set, and to the empty set in the case of the singleton of the universal class.

If $\{z : \phi\}$ simplifies to a form $\{z : u \in v\}$ where no more than one of u and v is z, there are cases to consider. If neither is z, then $\{z : (\forall w : \phi)\}$ reduces to the universe or the empty set depending on the truth value of $(\forall w : \phi)$. If neither is w, then $\{z : (\forall w : \phi)\}$ reduces to $\{z : u \in v\}$. $\{z : (\forall w : z \in w)\}$ simplifies to the empty set. $\{z : (\forall w : w \in z)\}$ is the singleton of the universal class.

Now we can verify that if P is a formula which contains exactly one free occurrence of z and we have $x = y$ and $P[z/x]$, then we can deduce $P[y/z]$: we have $P[x/z]$, so $x \in \{z : P\}$ so (by the definition of equality) $y \in \{z : P\}$, so $P[y/z]$, and the uses of comprehension are justified (though P is not necessarily stratified!). This justifies the full rule of substitution, because we can use this rule n times to get $P[y/z]$ from $P[x/z]$, if there are n free occurrences of z in P.

We fault Quine for not demonstrating this. It is *far* from obvious and we suspect him of having the bogus argument in mind.

In modern presentations of NF, equality is primitive and its rules are among the axioms.

We have a comment overall. It should be noted that though the historical role of this paper was to produce a weird new set theory which was either

[2]A reader asked about this, so it is useful to point out here that it is a theorem that $\{x : \phi\} = \{x : \psi\}$ holds if ϕ and ψ are equivalent: if they have the intended extension they are equal by extensionality, and if they do not, they are both the universal class.

a Problem (for some) or a Philosophical Cause (for others), this is probably not its actual intention. Reading it as a whole, I think the intention of the paper is to make a case for logicism, and to make the machinery supporting this case as simple as possible.

The note which asks whether the condition on $\in$-chains can be weakened to the assertion that they are acyclic in the sense that $x_1 \neq x_n$ if $n \neq 1$ can be answered in the negative. The class of pairs $(x, \{x\})$ admits a definition which is acyclic in this sense, for example, and it cannot be a set. However, much later, it has been shown that the criterion that the graph on variables whose edges are determined by membership statements in ϕ is acyclic in the usual sense of graph theory defines a consistent comprehension scheme which is actually equivalent to stratified comprehension (though not all stratified formulas are acyclic in this sense). Notice that an acyclic formula in the latter sense is certainly stratified, since there is at most one $\in$-chain from any variable to any other.