Formalizing the Metatheory of Logical Calculi and Automatic Provers in Isabelle/HOL
(Invited Paper)

Jasmin Christian Blanchette
Department of Computer Science
Vrije Universiteit Amsterdam
Amsterdam, the Netherlands
j.c.blanchette@vu.nl

Abstract
IsaFoL (Isabelle Formalization of Logic) is an undertaking that aims at developing formal theories about logics, proof systems, and automatic provers, using Isabelle/HOL. At the heart of the project is the conviction that proof assistants have become mature enough to actually help researchers in automated reasoning when they develop new calculi and tools. In this paper, I describe and reflect on three verification subprojects to which I contributed: a first-order resolution prover, an imperative SAT solver, and generalized term orders for \( \lambda \)-free higher-order logic.

1 Introduction
At programming language conferences such as POPL and ICFP, submissions are often accompanied by formalizations. Proof assistants are even used in the classroom to teach language semantics and type systems [69, 79]. Paradoxically, the automated reasoning community has largely stood on the sidelines of these developments.

Like the shoemaker’s children who go barefoot, we reflexively turn to “pen and paper”—by which we usually mean \( \LaTeX \)—to define our logics, specify our proof systems, and establish their soundness and completeness. The automatic/interactive divide of our community is part of the reason. Few automatic prover developers have first-hand experience with a proof assistant. Nevertheless, it stands to reason that the members of the automated reasoning community should be early adopters of proof assistants. If we cannot convince these close colleagues of the value of tools called “theorem provers,” how are we going to seduce mainstream mathematicians, philosophers, and engineers?

The IsaFoL (Isabelle Formalization of Logic) effort\(^1\) aims at changing this situation. This “coalition of the willing” was inaugurated in 2015, initially as a Bitbucket repository that would enable Mathias Fleury in Saarbrücken and Anders Schlichtkrull in Copenhagen to carry out their respective Ph.D. projects while avoiding duplicated work. Their foresighted and well-funded bosses, Christoph Weidenbach and Jørgen Villadsen, have made this project possible.

My motto for the project is “Coq at POPL, why not Isabelle at CADE?” But in fact, Isabelle has been represented at CADE and IJCAR for several years, thanks to Tobias Nipkow, Lawrence Paulson, Christian Urban, and others. Moreover, René Thiemann, Christian Sternagel, and their colleagues have been using Isabelle to formalize term rewriting for over a decade. Their IsaFoR library directly inspired IsaFoL.\(^2\)

Our main objective with IsaFoL is to develop libraries and a methodology to support modern research in automated reasoning, especially about propositional and first-order logic. Proof assistants can help when developing proofs, but also when modifying them, to study generalizations and variants. Reviewing becomes much easier when a formalization exists; reviewers can then focus on checking the definitions and theorem statements. Although my primary interest is in metatheory per se, once we have formal libraries, we can build verified provers and proof checkers on top of them.

The project has grown to include contributions from researchers and students in Amsterdam, Copenhagen, Gothenburg, Grenoble, Munich, Saarbrücken, and Zurich. Contributors include Heiko Becker, Alexander Bentkamp, Andreas Halkjær From, Alexander Birch Jensen, Peter Lammich, John Bruntse Larsen, Julius Michaelis, Tobias Nipkow, Nicolas Peltier, Andrei Popescu, Simon Robillard, Sophie Tourret, Dmitriy Traytel, and Petar Vukmirović. The IsaFoL repository on Bitbucket is where much of the development takes place. Once individual entries have become mature enough, they tend to migrate to Isabelle’s Archive of Formal Proofs, which is continuously tested and updated to work with the most recent version of Isabelle. The maintenance burden largely falls onto the Isabelle developers, following the “you break it, you fix it” principle.

The use of Isabelle/HOL was initially motivated by personal preference, but it has turned out to be a fortunate choice. Isar proofs [101], locales [4], (co)datatypes [9], the

\(1\)https://bitbucket.org/isafol/isafol

\(2\) Font aficionados will notice the divergent preferences concerning serifs and kernings. Pronunciation is also crucial: to avoid confusion, Japanophiles should clearly articulate \textit{izaforu} (IsaFoL) and \textit{izafo} (IsaFoR).
code generator [37], the Refinement Framework [52], Imperative HOL [25], and the Sledgehammer [75] and Nitpick [15] tools have all played important roles. Higher-order logic’s lack of expressiveness is often cited as a drawback of the system, but it has not been an issue for me. Using Isabelle has been an opportunity to practice what I preach, gaining insights into how the subsystems I am responsible for—the (co)datatypes, Sledgehammer, and Nitpick—could be improved.

In his invited SAS 2018 paper [72], Peter O’Hearn put forward the notion that

in most research papers one gets a description of where a project got to, but not how it got there. We get the results of science and/or engineering, but not a picture of how it developed. For the practice of doing science and engineering I’ve always thought the issue of “how it’s done” is an important one, and that examples describing how can have some value.

In this spirit, I describe three subprojects of IsaFoL to which I have contributed: a first-order prover based on ordered resolution (Section 2), an optimized SAT solver based on conflict-driven clause learning (Section 3), and generalized term orders for \( \lambda \)-free higher-order logic (Section 4). I reflect on the motivation and rewards of such work, both scientific and metascientific, and use this opportunity to survey related work, including the other IsaFoL entries (Section 5).

All along, my primary metascientific motivation has been to continue carrying out research at the intersection of interactive and automatic proving, enabling me to remain in the CADE orbit. After the Sledgehammer success story—an automatic prover at the service of interactive ones [75]—I was stimulated by the thought of reversing the formula. There is an undeniable self-referential thrill to the project. Formalize logic in logic. Employ automatic provers to formalize their own metatheory. But as the project advanced, I started to appreciate its deeper value. The overwhelmingly positive response at CADE, CPP, IJCAR, and JAR has convinced me that this research program was overdue.

I am excited about recent developments, where flaws and limitations revealed in formalization motivate new research. We are gradually moving from carrying out case studies, where the proof assistant aspect is predominant, to writing metatheory papers that only briefly mention the formal development in their introduction.

## 2 A First-Order Ordered Resolution Prover

Nearly two decades after its publication, the Handbook of Automated Reasoning [83] remains an invaluable resource for researchers in the area. In 2014, I started formalizing its Chapter 2, written by Bachmair and Ganzinger [3]. The first four sections present the metatheory of ordered resolution with selection, culminating in Section 4.3 with a refutationally complete first-order prover, RP, based on this calculus.

At the time, I had little experience using Isabelle, so it was fortunate that Traylor agreed to act as my mentor.

The chapter had been haunting me as a tsudokou for a couple of years. Formalizing it meant I would finally have to take it from my reading pile and read it thoroughly. But there were also sound scientific reasons to choose this target for formalization, as we remarked later [88, Section 1]:

The text is a standard introduction to superposition-like calculi (together with Handbook Chapters 7 and 27). It offers perhaps the most detailed treatment of the lifting of a resolution-style calculus’s static completeness to a saturation prover’s dynamic completeness. It introduces a considerable amount of general infrastructure, including different types of inference systems (sound, reductive, counterexample-reducing, etc.), theorem proving processes, and an abstract notion of redundancy. The resolution calculus, extended with a term order and literal selection, captures most of the insights underlying ordered paramodulation and superposition, but with a simple notion of model.

Traylor and I made quick progress in two intense weeks, reaching the crucial Section 4.3. This is where the resolution calculus is lifted from ground (propositional) to nonground (first-order) clausal logic and where the RP prover is introduced and shown to be refutationally complete.

At the ground level, ordered resolution consists of the single \((n + 1)\)-ary inference rule

\[
\frac{(C_i \lor A_1 \lor \cdots \lor A_n)^n \rightarrow A_1 \lor \cdots \lor \neg A_n \lor D}{C_i \lor \cdots \lor C_n \lor D}
\]

with multiple side conditions that restrict the search space. This rule is refutationally complete, meaning that any unsatisfiable set that is closed under applications of the rule will contain the empty clause \( \bot \). A redundancy criterion identifies clauses and inferences that can safely be ignored; for example, the unit clause \( p(a) \lor q(b) \lor a \lor b \lor \cdots \).

Next, Bachmair and Ganzinger introduce the concept of a theorem proving process: a transition system that starts with an initial clause set \( N_0 \) and where each transition \( \triangleright \) corresponds either to an inference or the removal of redundant clauses. Under a fairness assumption, the calculus’s completeness theorem characterizes the limit of a derivation \( N_0 \triangleright N_1 \triangleright N_2 \triangleright \cdots \).

Section 4.3 is where the trouble starts. For nonground clauses, the resolution rule takes the form

\[
\frac{(C_i \lor A_{i1} \lor \cdots \lor A_{ik})^n \rightarrow A_1 \lor \cdots \lor \neg A_n \lor D}{(C_1 \lor \cdots \lor C_n \lor D) \sigma}
\]

where \( \sigma \) is the most general unifier of the constraints \( A_{i1} \triangleright \cdots \triangleright A_{ik} \triangleright A_i \). Ordering restrictions block inferences where \( A_i \sigma \) is smaller than \( C \sigma \) or \( D \sigma \). A literal selection mechanism further prunes the search space.

Traylor and I initially stopped here. By a stroke of good fortune, Schlichtkrull decided to resume the proof two years
Formalizing the Metatheory of Calculi and Provers

...after months of labor, and with expert help from Waldmann, he reached the final \textbf{qed}. The resulting \textbf{J}ICAR 2018 paper \cite{holder} was well received by the anonymous reviewers, reaping a “strong accept” and two “accepts.” One of the reviewers wrote:

“The authors convinced me that their development is a great tool for exploring/developing calculus extensions. It will enable us to ‘extend/hack without fear.’”

(The italics are mine throughout this paper.)

The \textbf{RP} prover is naturally formulated in Isabelle as an inductive predicate. Bachmair and Ganzinger’s transition rules correspond directly to introduction rules:

\begin{verbatim}
inductive ~ :: 'a state ⇒ 'a state ⇒ bool where
\end{verbatim}

We faced various difficulties when formalizing Section 4.3. It did not help that the text contains dozens of small mistakes. Even the statement of the nonground resolution rule suffers from typos and ambiguities. While we agree with the reviewer who wrote that “most of us unconsciously auto-correct it, and read it with the intended meaning,” on several occasions we found ourselves blindly trusting the text, only to be disappointed later.

Reasoning about the \((n + 1)\)-premise resolution rule was particularly tedious. Ellipsis is a convenient pen-and-paper device that lacks a counterpart in proof assistants. We ended up keeping the clauses \(C_i\) and atoms \(A_{i,k}\) in parallel lists. Ideally, the first \(n\) premises of the rule would be regarded as a multiset; there is no need to consider \(n!\) inferences all yielding the same clause. However, there is no such things as “parallel multisets,” and our attempts at storing tuples \((C_i, A_{i,k})\) in multisets only made matters worse.

Another general difficulty with the chapter was that it is not always clear which hypotheses apply where. The text both presents a general framework and applies it, but dependencies are not tracked precisely, and many lemmas are never invoked explicitly. It was a challenge to understand the informal proof well enough to organize the locales and state the definitions and lemmas precisely and correctly.

Finally, some of the arguments are incomplete or misleading. The proof of Lemma 4.11 relies on the observation that “\(D'\) cannot be deleted by backward reduction or backward subsumption.” However, in principle, \(D'\) could be deleted due to the presence of a more general clause \(D''\), which in turn could be deleted due to \(D'''\), and so on. The key missing observation is that this process can be iterated at most finitely many times, generalization being a well-founded relation.

With the locales, definitions, and lemmas in place, carrying out the proofs was mostly straightforward. We relied heavily on Sledgehammer to discharge the proof obligations.

We did find a significant mistake in the chapter. Theorem 4.13 states that \(\text{RP}\) is refutationally complete, but this does not hold due to the improper treatment of inferences containing duplicate premises. Remarkably, we discovered the mistake several months after reaching the final \textbf{qed}, while reviewing our definitions. We had inadvertently “auto-corrected” Bachmair and Ganzinger’s definition of \(\text{RP}\).

Formalization helps track assumptions and dependencies precisely. It helps us answer questions such as, “Is Lemma 3.13 actually needed, and if so, where?” Indeed, such a question recently arose in the course of Bentkamp’s research on superposition \cite{bentkamp}. He wanted to understand why the literal selection function \(S_M\) is defined so that “(ii) \(S_M(C) = S(C)\), if \(C\) is not in \(K\).” He quickly got two replies. Waldmann wrote:

> As far as I can see, \(S_M\) is really only needed for ground instances, and then case (ii) is irrelevant. I guess they just wanted to define \(S_M\) as a total function.

Thanks to Isabelle, Schlichtkrull could be more confident:

> I tried to change the definition in the formalization to return the empty multiset if \(C\) is not in \(K\). With this definition the above properties also hold, and the proof goes through.

This anecdote nicely illustrates how formal proofs help generate knowledge and understanding. Here, they helped Bentkamp “extend/hack without fear.”

The second half of Bachmair and Ganzinger’s chapter, starting with Section 5, focuses on variations, such as non-standard clauses and hyperresolution. Most of these are not implemented in modern provers, and we did not attempt to formalize this material. Instead, Schlichtkrull, Traytel, and I further refined the abstract prover \(\text{RP}\) to obtain an executable functional program. This work is described in a manuscript \cite{manuscript} that was originally written for an audience interested in the principles of programming languages.

We started by defining the inductive predicate \(\text{RP}_\omega\), which resembles \(\text{RP}\) but adds a timestamp to clauses and a weight function. Bachmair and Ganzinger \cite[p. 44]{bachmair} mention this idea in a footnote, but they require the weight function to be monotone in both the timestamp and the clause size, claiming that this is necessary to ensure fairness. Although it often
makes sense to prefer smaller clauses, our proof reveals that this is not necessary to ensure fairness.

Next, we defined $\text{RP}_d$ as a deterministic functional program. $\text{RP}_d$ simply calls the auxiliary function $\text{RP}_d\_\text{step}$ repeatedly until a final state (with $N = P = \emptyset$) is reached. $\text{RP}_d\_\text{step}$ is a function of about 40 lines of code that is loosely modeled after Vampire’s main loop [98]. To introduce this possibly nonterminating function in Isabelle, we defined $\text{RP}_d$ by means of an option monad, using the \texttt{partial_function} command [48], so that it returns a value of the form Some $R$ if the computation terminates and None otherwise.

Finally, we made $\text{RP}_d$ executable by connecting it to IsaFoR, which provides first-order terms and operations on them, such as unification and the Knuth–Bendix order. We invoked Isabelle’s code generator [37] to export the prover to Standard ML. The resulting program, $\text{RP}_e$, consists of about 1000 lines of functional ML code, including dependencies.

After working hard to obtain an executable prover, we evaluated it on a representative subset of 1000 TPTP benchmarks against two leading provers, E and Vampire, as well as Metis [44], which is written in Standard ML. The table below gives the number of problems solved (proved or disproved) by each prover [87]:

<table>
<thead>
<tr>
<th></th>
<th>Vampire</th>
<th>E</th>
<th>Metis</th>
<th>$\text{RP}_e$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>833</td>
<td>770</td>
<td>527</td>
<td>353</td>
</tr>
</tbody>
</table>

Although our prover cannot yet challenge the state of the art, its performance is respectable and could be improved further using refinement. In his presentations, Natarajan Shankar often stresses the view that

\[
\text{algorithm} = \text{inference} + \text{strategy} + \text{indexing}
\]

Indeed, the performance of an automatic proof procedure comes largely from three sources: the calculus, the heuristics, and the data structures. $\text{RP}_e$ implements an excellent calculus, but mediocre heuristics and data structures. In the next section, we will look at a verification project that takes these two aspects seriously.

The work on formalizing $\text{RP}$ and $\text{RP}_e$ opens exciting perspectives. First, it paves the way for the formalization of superposition provers. Pelletier [77] took us by surprise when he announced, in 2016, his Isabelle formalization of the superposition calculus. Based on his work, it should be possible to define an SP prover analogous to $\text{RP}$, implementing most of the simplification rules and heuristics described in, for example, the E paper [90]. Verifying data structures such as discrimination trees, feature-vector indices, and fingerprint indices would pose interesting challenges. In unpublished work, Vukmirović has verified the metatheory of fingerprints for $\lambda$-free higher-order logic [99] using Isabelle.

Another perspective is to improve Bachmair and Ganzinger’s framework. In unpublished work, Waldmann, Tourret, Robillard, and I have conceived, with pen and paper, a framework that captures abstractly the lifting from a ground calculus’s completeness result to a nonground RP-like prover. An Isabelle formalization is underway, which should culminate with a streamlined proof of Bachmair and Ganzinger’s Theorem 4.13. The framework will also support infinitary inferences; these are useful for automating higher-order logic, where the unification procedure may yield an infinite stream of incomparable unifiers.

Around 2012, Vampire was extended with a SAT solver, using a novel architecture called AVATAR [98]. The empirical results were sensational, but a fundamental question was left unanswered: is AVATAR refutationally complete? The literature contains contradictory, erroneous definitions of the architecture [11, 80, 98], but following discussions with Giles Regier and his colleagues, I believe I have reached a precise definition, for which refutational completeness holds, while isolating a few potential sources of incompleteness in Vampire. Furthermore, Vukmirović and I have shown that labeled splitting [31], as implemented in SPASS, can be seen as an instance of a slightly generalized AVATAR. I hope to establish all of this formally in Isabelle, using the new framework by Waldmann et al. described above.

### 3 A CDCL SAT Solver

As part of a 2015 M.Sc. internship in Saarbrücken, Fleury started formalizing Weidenbach’s textbook draft, tentatively called Automated Reasoning—the Art of Generic Problem Solving. He continued as a Ph.D. student, focusing largely on SAT solving and the conflict-driven clause learning (CDCL) calculus implemented in most modern SAT solvers.

Inconveniently for us, there already existed several verified CDCL-based solvers [58, 60, 61, 71, 92]. We found a niche by emphasizing the stepwise refinement methodology and the connection between calculi variants, and by considering some aspects that had not been the focus of formalization before: clause forgetting, solver restarts, and incremental solving. We hoped this would suffice to get our submission accepted at IJCAR 2016; little did we expect to receive the best paper award. In the jury’s words, our paper [14]

formalizes a modern SAT solver via a chain of refinements in a proof assistant, contributing to the program of formalizing highly-technical research in the field of automated reasoning using tools developed in this field.

We considered both the abstract CDCL calculus described by Nieuwenhuis, Oliveras, and Tinelli [68] and a more concrete, implementation-oriented calculus proposed by Weidenbach [100]. The calculi are represented by inductive predicates on state tuples, roughly along the lines of the RP prover described in the previous section.

The proofs are largely elementary, relying on basic results about multisets and well-founded relations. We generally found Nieuwenhuis et al.’s arguments easy to follow, with the notable exception of a gap in their termination proof. Weidenbach’s concise proofs were often more challenging
to formalize. To give a flavor of his text, I quote a passage
that argues why it is impossible for the same clause to be
derived, or “learned,” twice:

By contradiction. Assume CDCL learns the same clause
twice, i.e., it reaches a state \((M, N, U, D \lor L)\) where \(L\)
is applicable and \(D \lor L \in N \lor U\). More precisely, the
state has the form \((K_n \cdots K_2 K_1 M_2 \cdots M_1, N, U, D \lor L)\)
where the \(K_i\), \(i > 1\) are propagated literals that do not
occur complemented in \(D\), as otherwise \(D\) cannot be of
level \(i\). Furthermore, one of the \(K_i\) is the complement
of \(L\). But now, because \(D \lor L\) is false in \(K_n \cdots K_2 K_1 M_2 \cdots M_1\)
and \(D \lor L \in N \lor U\) instead of deciding \(K_1\) the
literal \(L\) should be propagated by a reasonable strat-
egy. A contradiction. Note that none of the \(K_i\) can be
annotated with \(D \lor L\).

Fleury needed over 700 lines of Isabelle to capture this para-
graph. By changing the argument, he was later able to reduce
the formal proof down to 250 lines. The key was to exploit
invariance: by first establishing a strong invariant on SAT
solver states (namely, that all unit propagations have been
performed before deciding a new literal), he could easily
perform the key step of the argument (“the literal \(L\) should
be propagated by a reasonable strategy”) without having to
refer to earlier states or transitions.

Although Fleury did not discover any significant flaw
in the metatheory of CDCL (which would have been most
surprising), he did find a critical mistake in an extension,
described in Weidenbach’s draft, of CDCL with the branch-
and-bound principle. He came up with a counterexample
that invalidates the main correctness theorem, whose proof
confused partial and total models. Interestingly, he later no-
ticed the same flaw in a SAT 2009 paper by Larrosa et al. [56];
the issue is silently repaired in a subsequent article [57].

As a proof of concept, Fleury implemented a deterministic
SAT solver and extracted functional Standard ML code from
it, preserving the formal guarantees established about CDCL:
soundness, completeness, and termination. The resulting
program was very inefficient; it could solve none of the 2009
SAT Competition problems in reasonable time.

This was good enough for me but not for Fleury. With
Lammich’s help, he proceeded to specify the two-watched-
literal (2WL) scheme and other imperative data structures
[65], gradually departing from Weidenbach’s draft. This work
is described in a JAR article [12] and a CPP 2018 paper [32].

The 2WL scheme makes it possible to efficiently identify
candidate clauses for unit propagation and conflict detection,
which are the core CDCL operations. In each clause, the
solver distinguishes two watched literals—literals that can
possibly influence their clauses’ truth value in the solver’s
current state. The solver maintains the “2WL invariant” for
each clause. Unfortunately, the literature is imprecise about
the nature of this invariant and about when it is required to
hold. Fleury quickly found himself studying MiniSat’s source
code [30], and eventually came up with a precise formulation.
He specified 2WL as an abstract calculus, following
Weidenbach’s draft, and proved that its transitions preserve
the 2WL invariant. To get an executable program, he refined
the calculus in several correctness-preserving steps. The re-
finement methodology enabled him to inherit invariants,
correctness, and termination from previous layers.

The next refinement step implements the rules of the 2WL
calculus in a more algorithmic fashion, using the nonde-
terministic programming language provided by Lammich’s
Refinement Framework [52]. The language is built on top of
a nondeterminism monad. To give a flavor of it, I present the
code of a function that implements four calculus rules:

definition PCU_{algo} :: 'a lit × 'a clause → state → state
where
PCU_{algo} LC S = do {
    let (M, N, U, D, NP, UP, WS, Q) = S;
    let (L, C) = LC;
    let \(L'\) ← RES \(\{L' \mid L' \in \text{watched} C - \{L\}\}\);
    if \(L' \in M\) then
        RETURN S
    else
        if \(\forall L \in \text{unwatched} C. \neg L \in M\)
            if \(\neg L' \in M\) then
                RETURN (M, N, U, C, NP, UP, WS, Q)
            else
                RETURN (L'^c, M, N, U, D, NP, UP, WS, Q)
        else do {
            \(K \leftarrow \text{RES} \{K \mid K \in \text{unwatched} C \land \neg K \notin M\}\);
            \((N', U') \leftarrow \text{RES} \{\langle N', U' \rangle \mid\}
            \quad \text{updateclazz} (N, U) C L K (\langle N', U' \rangle)\);
            \(\text{RETURN} (M, N', U', D, NP, UP, WS, Q)\)
        }
}

The subsequent refinement steps optimize the data struc-
tures and specify heuristics that reduce the nondeterminism,
following ideas from the SAT literature and actual imple-
mentations. Multisets give way to lists, and clauses are now
expressed using Imperative HOL’s heap monad [25]. Using
Isabelle’s code generator [37], we can extract a self-contained
program in imperative Standard ML.

Since the CPP 2018 publication, Fleury has improved Isa-
SAT further by implementing four optimizations: restarts,
forgetting, blocking literals, and machine integers.
**Restarts** is a technique that enables the solver to explore another part of the search space. We can keep completeness by gradually increasing the interval between restarts. Our calculi included an optional Restart rule all along to allow such behavior, but it was not implemented in the first IsaSAT.

**Forgetting** removes some learned clauses—consequences of the initial problem clauses that are derived during solving. The abstract CDCL calculus included a Forget rule from the start. The main difficulty is that each clause now needs to store a Boolean flag indicating whether it is deleted. This requires adding a header to the clause data structure for storing this and other useful information that guides the heuristics, and adapting all the refinement proofs.

**Blocking literals** are literals stored directly in the watch list, next to a pointer (or index) to the clause, that can be checked directly without dereferencing the pointer. The information is redundant, but it often saves a pointer dereference. **Machine (64-bit) integers** are large enough to store clause indices and other numbers for all SAT problems arising in practice, and they are more efficient than ML’s unbounded integers. To use them without losing the formal guarantees about the program, we generate two versions of the prover’s body and connect them with code of the form

```latex
while ¬ done ∧ notOverflow do
  ⟨invoke the 64-bit version of the solver's body⟩;
if ¬ done then
  ⟨convert the state from 64-bit to unbounded integers⟩;
while ¬ done do
  ⟨invoke the unbounded version of the solver's body⟩
```

In a preliminary evaluation, Fleury ran IsaSAT against four other solvers on a collection of 3313 benchmark problems, consisting of all the SAT Competition problems from the 2009, 2011, 2013, 2014, 2016, and 2017 editions of the SAT Competition and the 2015 edition of the SAT-Race. The solvers Glucose [1] and CaDiCaL [10] represent the state of the art; MiniSat [30] is a well-known reference solver; and versat [71] is the only other efficient verified solver we know of. Since IsaSAT does not implement preprocessing techniques yet, CryptoMiniSat [93] was used to simplify all problems before benchmarking. The tests were run with a time limit of 30 minutes per problem. The table below shows the number of solved problems, whether satisfiable or unsatisfiable, for each system:

<table>
<thead>
<tr>
<th>Glucose</th>
<th>CaDiCaL</th>
<th>MiniSat</th>
<th>IsaSAT</th>
<th>versat</th>
</tr>
</thead>
<tbody>
<tr>
<td>1670</td>
<td>1645</td>
<td>1361</td>
<td>704</td>
<td>357</td>
</tr>
</tbody>
</table>

Given that competitive SAT solvers are extremely optimized programs, these results are very encouraging. However, we must bear in mind that it took two years of hard labor, and tens of thousands of Isabelle lines, to get from 0 to 704. An obvious avenue for future work would be to add optimizations such as pre- and inprocessing [45].

A benefit of having a verified SAT solver is that it can be employed as a backend in other verified tools. We have started looking into extending IsaSAT with theory reasoning, as in an SMT (satisfiability-modulo-theories) solver, with the goal of integrating it in CeTA [23], a verified safety and termination proof checker developed as part of IsaFoR.

## 4 Lambda-Free Higher-Order Terms Orders

The last subproject has a different flavor from the other two, in that the formalization arose as a side effect of carrying out research in automated reasoning and not as an end in itself. The Matryoshka project, which started in 2017 and funds a collaboration between Amsterdam, Nancy, Saarbrücken, and Stuttgart, aims at extending superposition provers and SMT solvers with higher-order features. As a stepping stone towards full higher-order logic, we started by focusing on a λ-free fragment. Unlike in first-order logic, variables may be applied, and function symbols may be given fewer arguments than they can take. This language is sometimes called “applicative first-order” in the term-rewriting community.

Bentkamp developed, under Waldmann’s and my supervision, a refutationally complete superposition calculus for λ-free higher-order logic and implemented it in a prototype prover developed with Cruanes [7]. He wrote his proofs directly in 	exttt{Htac}, which was possible only because he is extremely rigorous and could count on two experts to check his proofs—namely, Waldmann and our colleague Tourret. In principle, he could have started from Peltier’s Isabelle formalization of superposition [77], but it seemed more difficult than working in 	exttt{Htac}, especially given that he was more familiar with Waldmann’s informal proof, which has a different structure from Peltier’s.

Superposition relies on a well-founded term order. Together with Becker, Waldmann, and Wand, I designed two (families of) orders that generalize the familiar Knuth–Bendix order (KBO) and recursive path order (RPO). Becker formalized most of KBO during an internship; I completed his work and proceeded with RPO. This research was presented at FoSSaCS 2017 [21] and CADE 2017 [5].

The term orders are comparatively simple mathematical objects that lend themselves well to mechanization. I worked directly in Isabelle to define the orders and state their desired properties, starting with the ground case. The following fragment corresponds to my first attempt at defining a lexicographic path order (LPO), a special case of RPO, on ground terms:

```isabelle
datatype 'c tm = F ('c tm list)

context
  fixes <c:: 'c ⇒ 'c ⇒ bool
```

---

3\text{http://satcompetition.org/}

4\text{http://matryoshka.gforge.inria.fr/}
assumes irreflp \( (<_c) \) and transp \( (<_c) \) and 
\[ f <_c g \lor g <_c f \lor f = g \]
begin

definition chcksubs :: \( ('c \text{ tm} \Rightarrow 'c \text{ tm} \Rightarrow \text{ bool}) \) where 
chcksubs \( R s t \leftrightarrow \) (case \( (s, t) \) of \( (F f s, F g t) \Rightarrow \) (\( \forall s'. \in s. R s' \land s' = t \) \land (ss \neq []) \land last ss = last ts \rightarrow 
\( R (F f (\text{butlast ss}) (F g (\text{butlast ts})))) \))

inductive \( \lll <_c 'c \text{ tm} \Rightarrow 'c \text{ tm} \Rightarrow \text{ bool} \) where 
\( s < t \lor s = t \Rightarrow s < F (ts @ [t])) \) 
\( | s = F f ss \Rightarrow t = F g ts \Rightarrow s < t \Rightarrow 
F f (ss @[u]) < F g (ts @[u]) \) 
\( | s = F f ss \Rightarrow t = F g ts \Rightarrow f <_c g \Rightarrow 
\) chcksubs \( (\langle \rangle) s t \Rightarrow s < t \) 
\( | s = F f ss \Rightarrow t = F g ts \Rightarrow \) chcksubs \( (\langle \rangle) s t \Rightarrow 
\) lexordp \( (\langle \rangle) ss ts \Rightarrow s < t \)

lemma irrefl: \( \neg s < s \) 
lemma trans: \( s < t \rightarrow t < u \Rightarrow s < u \) 
lemma total: \( s < t \lor t < s \lor s = t \) 
lemma compat_fun: 
\( s < t \Rightarrow F f (ss @ s \cdot ss') < F f (ss @ t \cdot ss') \) 
lemma compat_arg: 
\( F f ss < F g ts \Rightarrow F f (ss @[s]) < F g (ts @[s]) \) 

end

The specification is very short and depends on no background theory beyond list operations (cons \(-\), append \@, butlast, lexordp). It is a perfect match for my counterexample generator Nitpick [15]. And indeed, the tool quickly finds a counterexample to the irrefl conjecture: given a type \('c\) with two distinct symbols \(f, g\) such that \(f <_c g\), we have \( f (g f) < f (g f)\). (Sadly, Nitpick’s rival Quickcheck [24] fails with the error “No type arity tm :: full_exhaustive.”)

Using Nitpick, I was able to converge to an almost correct design. A major flaw had escaped my attention, namely: without arities or typing constraints, LPO is not well founded, because it allows infinite descending chains such as 
\[ f b > f a b > f a a b > f a a a b > \cdots \]

This was noticed early on by Waldmann. Nitpick is helpless here, because it is based on finite model finding. It can be used to detect cycles but not acyclic divergence.

For the proofs, we drew our inspiration mostly from Baader and Nipkow’s textbook [2]. However, they cover only LPO and not RPO. When formalizing RPO, I faced a chicken-and-egg problem that took me several days to untangle. The issue is related to the multiset order, which is used to define RPO.

There exist two main formulations of the multiset order: Dershowitz–Manna [29] and Huet–Oppen [42]. They are equivalent on partial orders. Since RPO is a partial order, at first I chose Huet–Oppen, which we had used for KBO; however, until we have proved irreflexivity and transitivity of RPO, we cannot assume it is a partial order, so the choice between the two multiset orders is crucial. Zantema [103] was well aware of this, but I came across his work too late. And regardless, I could not think clearly while under the charm of the myth “Dershowitz–Manna = Huet–Oppen.”

As is often the case, once the main ideas were clarified, the formal proofs were straightforward to develop. According to generated logs, on some days I invoked Sledgehammer over 100 times. The first-order nature of most proof obligations (with the notable exception of well-foundedness) was a good match for automatic provers. The following Isar fragment, where each subproof (highlighted in gray) was produced by Sledgehammer, is fairly representative:

\[
\begin{align*}
\text{have} & \quad \text{arity_hd} \ (\text{head s}) = 1 \\
& \quad \text{by (metis One_nat_def arity.wary_AppE)} \\
& \quad \text{dual_order.order_iff_strict enat_defs(1), ileI1 linorder_not_le linorder_not_less} \\
& \quad \text{wary_st wt_gt_δ_if_supernumerary wt_s)} \\
\text{hence} & \quad \text{nargs_s : num_args s = 0} \\
& \quad \text{by (metis enat_or_simps(2) less_one nargsLt one_enat_def)} \\
\text{have} & \quad s = \text{Hd} \ (\text{head s}) \\
& \quad \text{by (simp add: Hd_head_id nargs_s)} \\
\text{then obtain} & \quad \text{f where} \\
& \quad f \in \text{ground_heads} \ (\text{head s}) \land \\
& \quad \text{wt_sym f + the_enat (δ * arity_sym f) = δ} \\
& \quad \text{using exists_wt_sym wt_s by fastforce}
\end{align*}
\]

The first proof above relies on 12 named lemmas or local facts. Developing such a proof manually could easily have taken half an hour or more without the help of Sledgehammer (to find the proof) or the metis proof method (to reconstruct it), and would have required searching for lemmas or memorizing their names.

In our two papers and the associated technical reports, we presented informal versions of the Isabelle proofs, for human consumption. This has been an opportunity to clean up and restructure the Isabelle proofs, to emphasize the important steps. We also used Nitpick to create examples to illustrate fine points in the papers [5, Example 12; 21, Example 9]. Given two orders \(<_1\) and \(<_2\), we could for example ask the tool to generate terms \(s, t\) such that \(s <_1 t\) but \(s >_2 t\).

Proof assistants really come into their own when we start modifying existing developments. Late in the project, we asked ourselves, “Could we generalize definition so-and-so in such-and-such a way?” After spending one hour with Isabelle, I was convinced the answer was yes, and a few hours later I was doing the proofs. There was very little to change in the technical report, but I would have had a hard time locating the passages that needed modifications and convincing myself that I had found them all.

The two papers received a lukewarm welcome. The term orders were not implemented in any termination tool, nor (at the time) in a superposition prover, and some related work
was not treated. Nevertheless, the anonymous reviewers were impressed by the formalization, which probably saved the submissions. I realize I may be preaching to the choir, but let me quote two of the reviews:

Statements are very precise. Only necessary conditions are made. In particular, orders are characterized by 9 properties, and for each main statement, the authors discuss which of them are necessary.

[The paper] incorporates a comprehensible structure, precise assumptions and thorough proofs for all the KBO’s essential properties. All the presented proofs (irreflexivity, transitivity, subterm property, compatibility with functions and arguments, stability under substitution, totality on ground terms, well-foundedness) have been formalized within Isabelle/HOL and published as part of the Archive of Formal Proofs. Together with the careful presentation of these proofs within the submission, this establishes a high overall reliability of the presented results.

Isabelle helped in other ways during the CADE rebuttal phase. A reviewer had asked, “Doesn’t X2 follow from X1 by taking \( x > y \) iff \( h(x) > h(y) \)?” Using Nitpick, we were able to provide a simple counterexample.

Our \( \lambda \)-free higher-order RPO suffers for a theoretical blemish: it does not satisfy compatibility with arguments, the requirement that \( s < t \) implies \( s < u < t \) for all terms \( s, t, u \). In fact, this requirement cannot be met while maintaining the subterm property and coincidence with the standard RPO on the first-order fragment of higher-order terms. In 2017, I challenged Bentkamp to design an RPO-like order that satisfies compatibility with arguments, by giving up full coincidence with the standard RPO. He succeeded brilliantly.

He even used Isabelle to design his new order, the embedding path order (EPO) [6], and reports that proving the properties using Isabelle was easier than using pen and paper. On the other hand, he struggled with my tool Nitpick and ended up testing his conjectures by coding them in Haskell and using Lazy SmallCheck [84].

At some point, I generalized our definition of KBO to use ordinals instead of natural numbers, resulting in a transfinite KBO (TKBO). For this work, I collaborated with Fleury and Traytel. Using Isabelle’s new (co)datatype package [9], we could define the syntactic ordinals—the ordinals below \( e_0 \), representable in Cantor normal form—as the following datatype of “hereditarily finite multisets”:

\[
\text{datatype} \quad \text{hnmultiset} = \text{HMSSet} (\text{hnmultiset} \ \text{multiset})
\]

The recursion through a nondaatatype—multiset—would be problematic in other systems, or in pre-2013 versions of Isabelle/HOL. Using this datatype, an ordinal \( \alpha = \omega^{	ext{c_1}} \cdot c_1 + \cdots + \omega^{	ext{c_n}} \cdot c_n \), where \( c_i \in \mathbb{N}^+ \) and \( \alpha_1 \succ \cdots \succ \alpha_n \), is represented by

\[
\alpha = \{ \alpha_1, \ldots, \alpha_1, \ldots, \alpha_n, \ldots, \alpha_n \}
\]

We used this opportunity to introduce a type of nested finite multisets and defined Dershowitz and Manna’s nested multiset order on it [29]:

\[
\text{datatype} \quad \text{a nnnmultiset = Elem a | MSet (a nnnmultiset multiset)}
\]

This enabled us to finally give a positive answer to Paulson, who in 2014 had asked on the Isabelle mailing list:

I wonder whether anybody is aware of a formalisation (in any system) of the nested multiset ordering, as described in the classic paper “Proving Termination With Multiset Orderings”?

Using Isabelle’s Lifting and Transfer tools [43], we established a bijection between \( \text{hnmultiset} \) and the Elem-free fragment of \( \text{a nnnmultiset} \) and exploited it to lift definitions and properties. Notably, lifting the nested multiset order gives the familiar \( < \) operator on ordinals. The order’s well-foundedness proof can be transferred as well. Ordinal arithmetic operations such as addition and multiplication can be defined directly in terms of multiset operations.

Overall, we were able to quickly develop a versatile, practical library of syntactic ordinals, which we used not only for our TKBO variant but also in a formal proof of Goodstein’s theorem. This research was presented at FSCD 2017 [13].

5 Related Work

IsaFoL consists of many more subprojects beyond those described above. The first two listed below predate IsaFoL, but they are very much in its spirit and are mentioned on its web page. The entries are listed in rough chronological order:

- equisatisfiability of sort encodings for first-order logic, by Popescu and myself [16, 17];
- abstract soundness and completeness results for first-order logics using coinductive methods, by Popescu, Traytel, and myself [18–20];
- soundness and refutational completeness of first-order unordered resolution, by Schlichtkrull [85, 86];
- soundness and refutational completeness of a generalization of the superposition calculus, by Peltier [77];
- soundness and completeness of resolution-based prime implicate generation, by Peltier [76];
- metatheoretical results about a paraconsistent propositional logic, by Schlichtkrull and Villadsen [89, 96];
- soundness of a small-kernel first-order prover with equality described in Harrison’s textbook [40], by Jensen, Larsen, Schlichtkrull, and Villadsen [46, 47];
- correctness of an optimized tool chain for checking SAT solver certificates, by Lammich [54, 55];
- various metatheoretical results about a wide range of proof systems for classical propositional logic (sequent calculus, natural deduction, Hilbert systems, and resolution), by Michaelis and Nipkow [63, 64];
Formalizing the Metatheory of Calculi and Provers

- extensions of Berghofer’s formalization [8] of a first-order natural deduction calculus, by From [33];
- soundness of a substitutionless proof system for first-order logic, by From, Larsen, Schlichtkrull, and Villadsen [35];
- soundness and completeness of an epistemic logic with countably many agents, by From [34];
- modernization of Ridge and Margetson’s formalization [81, 82] of a sequent calculus for a term-free first-order logic, by Villadsen, Schlichtkrull, and From [97].

Formalizing metatheoretical results about logic, proof systems, and reasoning tools is an attractive proposition for many researchers in our field. Landmark achievements in the 1980s and 1990s include Shankar’s proof of Gödel’s first incompleteness theorem in Nqthm [91], Paulson’s completeness proof for intuitionistic predicate logic in ALF [78], and Harrison’s formalization of basic first-order model theory in HOL Light [38].

Following Shankar’s 1984 proof, Gödel’s first incompleteness theorem has been formalized in Coq by O’Connor [70], in HOL Light by Harrison (in unpublished work), and in Isabelle/HOL by Paulson [73, 74]. Paulson also succeeded at verifying the second incompleteness theorem.

The completeness theorem for first-order logic has been mechanized many times in proof assistants. In Isabelle/HOL, Berghofer [8] proved the completeness of a natural deduction calculus, and Margetson and Ridge [81, 82] proved soundness, completeness, and cut-elimination of a sequent calculus for a term-free first-order logic. I refer to a recent article I wrote with Popescu and Traytel [20] for a discussion of such work.

Term rewriting is another popular target of formalization. The CoLoR library by Blanqui and Koprowski [22] and the CiME3 toolkit by Contejean et al. [26], both in Coq, as well as IsAfoR [95] in Isabelle, have explored this territory. They include not only formalized metatheory but also verified (non)termination and (non)confluence checkers built on it.

SAT solving also lends itself particularly well to formalization. Marić [59, 60] verified a CDCL-based SAT solver in Isabelle/HOL, including watched literals, as a purely functional program. He also formalized the abstract CDCL calculus by Nieuwenhuis et al. and, together with Janičić [61], the more implementation-oriented calculus by Krstić and Goel [49]. As a milestone towards verified SMT solvers, Spasić and Marić [94] formalized the simplex algorithm in Isabelle; Thiemann extended this work to support incremental solving and to provide unsatisfiable cores [62]. In Coq, Lescuyer [58] formalized the CDCL calculus and the core of an SMT solver. Another verification of a CDCL-based SAT solver, including termination but excluding two watched literals, is by Shankar and Vaucher [92] in PVS. Impressively, most of this work was carried out by Vaucher during a brief internship. Finally, Oe et al. [71] used Guru to specify and verify the SAT. The generated C program consists of 15,000 lines of code. Optimized data structures are implemented, notably for watched literals. However, termination is not guaranteed, and model soundness is established through a run-time check.

A pragmatic approach to obtain trustworthy unsatisfiability judgments from a SAT solver is to have it produce a certificate, which can be given to an independent checker. An efficient format for this is DRAT [41]. The standard DRAT checker [102] is an unverified C program, but verified checkers have now been developed by Cruz-Filipe et al. [27] and Lammich [54] using ACL2, Coq, and Isabelle/HOL.

I alluded, in the introduction, to the self-referential thrill of formalizing theorem provers. Harrison [39] took this idea to its logical extreme when he verified, in HOL Light, HOL Light’s inference kernel. To circumvent the impossibility of defining higher-order logic’s semantics in itself, he carried out two distinct proofs: one where the formalized logic has no infinity axiom, and one where HOL Light is extended with an axiom to increase its strength. Harrison’s formalization was ported to HOL4 and extended by Kumar et al. [50] to include definitional mechanisms and to exploit CakeML [51], a verified ML environment. In another line of work, Davis built an ACL2-style prover called Milawa [28]. The development consists of a stack of provers, each used to verify the one above it. Together with Davis, Myreen [67] connected Milawa to a verified Lisp implementation [66] that was developed for hosting Milawa. A noteworthy feature of the prover is its switch command, which lets the user replace the inference kernel by an arbitrary kernel that has been proved sound, enabling powerful, highly optimized extensions that would be impossible using a traditional LCF architecture [36].

6 Conclusion

In this paper, I reported on some of the steps my colleagues and I have taken to help drive the adoption of proof assistants in the automated reasoning community. Far from following a definite plan, at every turn we focused on topics that appealed to us and for which we could perceive clear value in formalization. We have barely scratched the surface; a lot of exciting work still awaits us.

Automated reasoning is near ideal territory for proof assistants. Compared with other application areas, the proof obligations are manageable, and little background theory needs to be formalized before we can get started. Conveniently, researchers in the area are not afraid of logic, although they often lack familiarity with proof assistants and their higher-order formalisms.

Isabelle/HOL has been a very suitable vehicle for this kind of work, and we will continue using it. It is comparatively easy to use, has a simple but expressive logic, is based on a trustworthy LCF-style foundation, and includes rich libraries developed by a large, and growing, user base.
Acknowledgments

I am grateful to all contributors to the IsaFoP effort, listed in the introduction, and to my colleagues Maximilian Kirchmeier, Christian Sternagel, Uwe Waldmann, Daniel Wand, and Christoph Weidenbach. This paper would not have existed without the dedicated labor of Heiko Becker, Alexander Bentkamp, Mathias Fleury, Peter Lammich, Anders Schlichtkrull, and Dmitriy Traytel; and the research program might never have started without the early support and advice of Tobias Nipkow, Andrei Popescu, Dmitriy Traytel, Jørgen Vilaldsen, and Christoph Weidenbach. Discussions with Franz Baader, Pascal Fontaine, Carsten Fuhs, Marijn Heule, Cynthia Kop, Leonardo de Moura, Michael Norrish, Lawrence Paulson, Giles Reger, Stephan Schulz, Natarajan Shankar, Cesare Tinelli, Andrei Voronkov, and many others have influenced this research. The Isabelle development team and the editors of the Archive of Formal Proofs have made the work possible. I thank Assia Mahboubi and Magnus Myreen for the invitation to write in these pages and for their useful comments. Alexander Bentkamp, Mathias Fleury, Pascal Fontaine, Robert Lewis, Anders Schlichtkrull, Mark Summerfield, Sophie Tourret, and Dmitriy Traytel have suggested many textual improvements. I should also acknowledge that some of the text above has appeared in a similar form in my earlier papers [12, 21, 32, 87, 88]. The research has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (grant agreement No. 713999, Matryoshka). My work was also partially supported by a Deutsche Forschungsgemeinschaft (DFG) project (grant Ni 491/14-1, Hardening the Hammer).

References

Formalizing the Metatheory of Calculi and Provers


1264

1264 6:1–6:16.


