Program Analysis Techniques
for Model Queries and Transformations

PhD thesis booklet

Zoltán Ujhelyi
MSc in Technical Informatics

Advisor:
Dr. Dániel Varró, DSc
Professor

Budapest, 2016
1 Preliminaries

Model-driven Engineering The discipline of model-driven engineering (MDE) [BG05; WHR14] aims to address the increasing complexity of software development by focusing on the definition and analysis of high level models that describe various aspects of the developed system on different levels of abstraction. This approach provides better understandability, faster development lifecycle and increased software quality, e.g. fewer bugs [HWR14], which makes MDE especially useful in critical system development where system failure may lead to human injury or significant damage in property [Bro+14].

Simultaneous increase in quality and productivity is primarily achieved in MDE by early validation of system models and automated generation of various design and implementation artifacts. Early validation allows to highlight conceptual flaws well before such flaws are detected by traditional testing techniques. Automated source code generation from high-level behavioural models reduces both development and verification time. Furthermore, high-level models also enable the automated synthesis of test cases, configuration files, documentation, etc.

As models become ubiquitous as the main artifacts in systems development in many application domains [Béz05], a wide range of modeling languages have been introduced. They may represent the software or hardware architecture (e.g., UML [Obj11c], SysML [Obj10] or AADL [SAE09]), the external or internal behavior (e.g., UML statecharts [Obj11c], BPMN [Obj11a], BPEL [OAS07]), the reliability or performance of components (e.g., MARTE [Obj11b]) or support formal analysis, such as Petri nets [Mur89] to model distributed system or Kripke structures used in model checking [Cla08]. Furthermore, particular industries often make use of platform description languages, such as AUTOSAR [AUT12] in the automotive domain, or ARINC 653 for Integrated Modular Avionics [Pri08].

Model Queries and Transformations Several important engineering tasks supported by model-based tools rely on model queries. Such queries, by definition, retrieve those elements from a model that satisfy a query condition, which can be a complex composition of subconditions. Their applications include early model analysis [Aut+16; Bar+15], like well-formedness constraint validation or the calculation of various model metrics. Moreover, model queries can to ease the derivation of higher-level view models.

Automated model transformations (MT) are used to create or update models in different modeling formalisms from one or more existing models. Model transformations may generate various artifacts, like test cases, documentation or other models [Bon+99; Rus08]. Furthermore, by the generation of formal models, it is possible to execute formal analysis on the developed models while effectively hiding the theoretical mathematical details and enabling the designers to use well-known, high-level models to specify their systems [Bon+99].

Code generators that map models to textual artifacts, like source code or configuration descriptors can be considered as a special kind of model transformation where the output is either directly text, or an abstract syntax graph of the textual output. Such generators, like Accceleo [Acc], the Epsilon Generation Language (EGL) [Epsilon; Ros+08] or Xtend [Xtend], are widely used in modeling tools, and are often considered an important aspect of productivity gain by MDE [HWR14].

The correctness of such model queries and transformations is critical as undetected errors can invalidate the analysis results or they might propagate into the target application. Due to the increasing complexity of queries and transformations, ensuring their correctness becomes increasingly difficult.
1.1 Analysis of Model Queries and Transformations

Model transformations are captured in the form of a model transformation program, which is a regular piece of software. Elementary model transformation steps are often formulated by highly data-driven, declarative transformation rules, while imperative control structures help assemble complex transformations from these elementary steps. Furthermore, the input and output models of transformations are complex data structures. As a consequence of this, existing validation results of software engineering has to be extended when adapting them for transformation programs [Bau+06; Küs06], especially for designing complicated, industrial transformations, where debugging and validation plays a crucial role.

There is a wide range of techniques available for the verification of model transformations, surveyed in [AW13]. Following its categorization, the approaches range from generating test cases and oracles, formal analysis using theorem proving, model checking to graph theory-based proofs.

- The main challenge of testing model transformations [Bau+10] is their complex, but structured input and output data (models), the use of model management environments (e.g. to create or validate models), and a large set of transformation languages and techniques. To generate instance models that cover the transformation programs sufficiently, often equivalence partitioning [Stu+07; WKC08] is used. In [FSB04] an alternative, mutation-based approach is proposed. To provide test oracles, multiple approaches are available [MBL08]: provide a known-to-be-correct reference transformation, and compare the results; provide an expected output model manually; or describe the expected results using constraints, e.g., by the use of OCL [Bau+06] or graph patterns [OW09].

- **Theorem proving** techniques are also very diverse. Direct approaches consider transformation rules directly, either by generating an equivalent of the transformation as the input of a theorem prover tool [Cal+11], or by extracting some invariants from the specifications [Cab+10; ALL09]. Other, indirect approaches verify properties of the output model, such as executing well-formedness analysis [ER10] or checking the conformance to semantics metamodel [Bar+08]. Further techniques, including the Amphion tool [Van+98] used by NASA, or the AutoBayes code generator [Sch+03], provide evidence to check its results. In general, indirect approaches are less complex as they only consider the outputs generated by the transformation, however, the transformation itself is not verified, and requires additional analysis after each execution.

- **Model checking** methods focus generally on verifying dynamic properties of the model transformations, such as termination using techniques like state space exploration [LBA10; Var04; Ren04] or bisimulation [KN07]. A mapping to Colored Petri nets was proposed in [Wim+09] to verify static and dynamic properties, including termination or liveness. The main benefit of model checker tools is that in case of verification failure, a counter-example is retrieved showing an execution trace to the location of the fault; on the other hand, the well-known state-space explosion problem frequently limits its practical usability. An additional problem in case of models is that the graphs themselves may grow (creation or deletion of new objects) resulting in infinite state-spaces [BCK08; RD06] requiring the introduction of some abstractions over the graph models.

- As models are usually expressed as graphs, the verification of transformation methods, especially graph transformations, can rely on mathematical proofs based on graph theory. Several transformation properties are already evaluated this way, such as termination [BKZ14; Var+06] or confluence [Küs06; TG15]. Additionally, as graph transformation
approaches can be used to specify bidirectional transformations, verification of invertibility [Her+11] is also useful.

When comparing the various techniques, a commonality is that their usage requires medium-to-high level effort [AW13], since the respective papers focus more on the core algorithms themselves rather than their automation and adaptation. Additionally, the sheer complexity of the analysis tasks make most verification approaches computationally intensive.

A complementary approach would be the use of lightweight analysis techniques, such as static type checking or (pattern-based) static analysis tools (like the FindBugs [Fin] or PMD [PMD14] tools for Java programs) to identify common developer problems during development. However, less research has been dedicated to investigate such techniques for model transformations. In fact, such techniques are omitted from [AW13] despite their practical usefulness is already proven in other software engineering applications.

1.2 Debugging Model Queries and Transformations

In traditional software engineering if a (permanent) software fault is detected in a program, developers begin a process called debugging [STJ83], where the main goal is to localize the fault. Finally, the localized fault is repaired.

Fault detection can either be a result of automated or manual testing, or it can be initiated by a report from the end-user. Generally, fault localization is considered the most complex part of debugging; localized faults are usually easy to fix. Debugger tools usually allow stopping the program execution at any given time using breakpoints, and the current program state, such as the set variables, is presented while allowing modifications as well.

Debugging transformation programs [MV11] may conceptually work similarly. Usually, faults are detected by finding incorrect output models, resulting in simple questions [HLR07] like "Why are there no objects of a selected type in the target model?" or "What might have caused this violation of a metamodel constraint?".

However, fault localization is problematic for similar reasons as testing model transformations: inputs and outputs are complex, structured models defined in model management environments. When halting a transformation program at a breakpoint, the current state must include the program variables and the models as well. Furthermore, in case of large models, context-sensitive filtering is required to visualize the models meaningfully.

Furthermore, in case of high level transformation languages, especially that of declarative nature and/or graphical syntax, there might be a large conceptual gap between transformation definition and execution. E.g., most transformation languages offer primitives to evaluate complex conditions, like OCL expressions or graph patterns. This is especially the case for translation-based approaches where a regular piece of software is derived from the transformation specification [Wag+11].

There are different approaches that address some issues of debugging model transformation programs, such as input model tainting [Dho+10] that mutates existing models to identify corner cases where the transformation is incorrect; or the use of omniscient or forensic debuggers [HLR07; Cor14] that take logged execution traces, and allow both forward and backward stepping during the executed steps.

Still, current transformation development environments provide only limited debugger support, requiring additional work for the remaining challenges such as integrated model visualization or filtering support.
2 Research Questions and Challenges

My research has been motivated by practical challenges to assist the lightweight analysis of model queries and transformations for developers by locating and highlighting errors. As opposed to complex verification approaches, my focus was to propose lightweight techniques that provide direct, integrated feedback to developers with a fast response time.

Such lightweight analysis techniques are often useful as they can be fast enough to report errors as soon as the developer commits them, making it easy to provide a fix. Similarly, during debugging, analysis techniques may provide a hint where the problem is potentially located. Lightweight techniques are less precise than formal verification approaches, thus they cannot guarantee that a query or transformation is free of flaws. However, their fast, integrated nature complements verification techniques by providing early error detection.

I identified three challenges for the lightweight analysis of model queries and transformations.

Challenge 3: How to observe the structure of models during execution?

During the debugging of model transformations it is common that certain parts of model elements are important with regards to the error being investigated, but no model visualization is available that directly displays a developer-specified subset of the models, and updates it as the transformation executes. Such a visualization would be useful to monitor the active state of an internal state machine or follow the creation of traceability links during the process of model transformation.

There is a large number of methods and tools for displaying graphical views of models [God+07; Bul08; WEK04]. However, they are not suitable for the definition of ad-hoc visualizations required by the transformation debugging scenario, as they either require manual customization to filter the model to display the right information, or require coding to create the visualization, making the approach more expensive to use.

Furthermore, they either require to export the model before they can be displayed, or require additional programming effort to display the model. Model export can usually be written easily - in some cases, such as in the GrGen.net tool [Gei+06] it is available with the framework, - however, to monitor the changes in the transformation this model needs to be updated. Manually integrating visualization tools to the transformation is a possibility, but for transformation debugging scenarios the implementation effort often exceeds the cost of finding the related error without the help of such a visualization.

Other environments like Sirius [Sirius] or Papyrus [Papyrus] allow defining view models, but they still require significant effort to develop (mainly because the complex feature set of the tools), and updating the visualization after changes requires careful planning.

In this thesis I propose a lightweight approach to define graphical views based on annotated graph patterns. The approach relies on incremental graph pattern matching and event-driven model transformations to provide an always current notation model that can be displayed using a wide range of model visualization tools.

Challenge 1: How to efficiently detect frequent flaws in model query and transformation specifications?

During model query and transformation development, it is easy to make mistakes that result in syntactically correct queries that always return an undesired result set. These problems are hard to find in a debugging session because of the high-level, declarative specification. On the other
hand, some common causes can be efficiently identified, such as incorrect ordering of parameters
during a call; referencing an incorrect variable or even some copy-paste errors, e.g. missing or
extra constraints in a query.

As these issues are usually trivial to fix once identified, automatic detection is most effective
if it can be executed when editing the queries or transformations. However, this requires an
efficient approach that does not require a high level of resources. Furthermore, it is required to
back-annotate the internal analysis results to the original source code, and present the results
visually.

Type checking and the verification of the graph structure of pattern definitions are two ways
to detect common errors statically: they can be executed fast, while still identifying common
problems mentioned before. In this thesis, I will propose to perform such analysis over model
queries and transformations and evaluate its capabilities.

**Challenge 2: How to identify the root cause of complex transformation errors?**

During debugging a developer often starts from the point in the program where an error first
manifests, and attempts to reason about the executed operations that led to that state. This kind
of reasoning is useful during model query or transformation debugging as well, however it must
include both the model query or transformation program and the model the it was executed on.

Such a process can be supported by automatically highlighting model elements and program
statements that might contribute to the transformation error, and the developer does not have to
consider the excluded (independent) elements and program points, making it easier to identify
the root cause of the issue.

Program slicing [Wei82] is a widely used technique in traditional programming languages [BGG06;
GL91; RH07; Tót+10; SGM02] for such root cause detection. In this dissertation, I define the slic-
ing problem for model queries and transformations, and propose two algorithms for calculating
such slices.

### 3 Novel Scientific Contributions

#### 3.1 Static Type Checking of Model Queries and Transformations

There is a class of common programmer errors, such as incorrectly ordered parameters or invalid
variable assignments, that can be efficiently revealed statically on model queries and transfor-
mations by executing type checking, usage counting or reachability analysis. As such issues
occur commonly, revealing them before executing helps query and transformation developers
providing error-free programs. I formulated the following contributions related to the detection
of erroneous specification (in fulfillment of Challenge 1):

**Contribution 1** I proposed a set of lightweight static analysis techniques to identify common
mistakes in model queries and transformations: type checking, variable usage counting and
reachability analysis. I integrated these approaches into the development environment of the
V/IA/T/RA framework.

1. **Type checking and type inference with constraint satisfaction problems.** I defined
type checking and type inference by mapping type safety of model queries and transfor-
mations to a series of constraint satisfaction problems. The proposed mapping provides
detailed localization in case of type errors [4, 15, 25, 26].
2. **Type constraints of model queries.** I defined type rules for graph patterns supporting all constraint types of the languages of the VIATRA and EMF-IncQuery frameworks [4, 14, 26].

3. **Type constraints of transformation programs.** I defined type rules for graph transformation rules and abstract state machine rules, and evaluated its performance using various traversal strategies [4, 15, 26].

4. **Static analysis techniques for model queries.** I proposed variable usage counting and reachability analysis to detect common errors in model queries [3], such as unused variables and disjunct constraints.

My results build upon the query and transformation languages of the VIATRA and EMF-IncQuery frameworks. The VIATRA transformation language [VB07] was developed by András Balogh and Dániel Varró, while the query language of EMF-IncQuery [14] is part of Gábor Bergmann’s thesis [Ber13].

### 3.2 Incremental Graphical Views

Queries defined for a transformation specify some aspects of the underlying model that is relevant during the debugging of transformation programs. Reusing these queries as lightweight declarations of visualizations allows fast creation of graphical views, making it possible to visualize interesting parts of the models fast. Furthermore, incremental query evaluation techniques would enable the implementation of efficient update mechanisms, visualizing the changes of the model, e.g. during a transformation.

I formulated the following thesis contributions for the observation of the structure of models during execution (fulfilling Challenge 2):

**Contribution 2** I defined an approach to derive and incrementally maintain graphical views defined declaratively by model queries. I demonstrated its feasibility using an implementation over the VIATRA framework.

1. **Graphical view definition using annotated queries.** I extended the language of EMF-IncQuery with annotations that allow the declarative definition of graphical view models. The annotations uniformly define how to display data in various (graphical, tree- or list-based) viewers [14, 16].

2. **Incremental graphical view maintenance based on incremental queries.** I proposed an approach for deriving and maintaining graphical views using incremental query evaluation techniques by mapping query results to displayed items and translating change notifications to the addition or removal of view elements. The mapping rules are automatically derived from the view definition.

The implementation of the graphical views builds on the incremental pattern matcher techniques [Ber+10] of EMF-IncQuery, which is part of the PhD thesis of Gábor Bergmann [Ber13],
and an event-driven virtual machine where Ábel Hegedűs was the main contributor. Csaba Debreceni has generalized my approach to support the construction of incrementally maintained arbitrary logical view models using a more generic set of annotations in our joint paper [16] together with a performance assessment; my own contributions to the work are the annotation-based view model definitions, the notational metamodel and the transformation workflow. In his MSc thesis, Péter Lunk (under my supervision) implemented a prototype debugger for ViATRA exploiting the annotation-based graphical views.

3.3 Slicing Model Transformations

Program slicing [Wei82] approaches are widely used for dependency analysis in traditional programming environments [Xu+05; Tip95]; similar approaches have been proposed in the context of MDD for model slicing [LK10; SP08; Sha+10]. However, the direct application of these techniques is problematic for model transformations, as it requires simultaneous slicing of the models and the transformation program. I formulated the following contributions to locate root causes of complex transformation errors (to address Challenge 3):

**Contribution 3** I defined the concepts of model transformation slicing to identify the dependencies between model elements and transformation program constructs. I adapted dynamic backward slicing algorithms for model transformation slicing and demonstrated the feasibility of the approach in the context of modern model query and transformation frameworks.

1. **Slicing of Model Transformations.** I defined the problem of slicing model transformation programs to identify model elements and transformation constructs dependent on a slicing criterion. I identified the main challenges of model transformation slicing by adapting traditional concepts of program slicing [12].

2. **Dynamic Backward Slicing of Model Queries.** I defined a technique that calculates dynamic backward slices during the evaluation of model queries based on Rete networks [6].

3. **Dynamic Backward Slicing of Model Transformation Programs.** I developed an algorithm that calculates dynamic backward slices of model transformation programs based on recorded execution traces of transformations [11].


4 Application of Results

The practical relevance of the methods and techniques outlined in the current thesis are demonstrated in this section by listing some recent applications of the scientific results.

**Tools I: EMF-IncQUERY Development Environment**

The EMF-IncQUERY open source model query framework adapts the Rete algorithm for incremental query evaluation techniques over EMF models, one of the most widely used modeling
environments today. Query evaluation serves as a basis for additional features such as automated validation, data binding, incremental graphical views and model transformation. The EMF-IncQuery framework is a major research contribution of Gábor Bergmann’s PhD thesis.

The experience gained when developing the type checker for VIATRA (Contribution 1) influenced both the design and the implementation of the pattern language of EMF-IncQuery: on one hand, inferrable type constraints can be omitted, resulting in shorter pattern definitions; on the other hand optional type definitions for pattern parameters can reduce the runtime cost of evaluating the type safety of pattern calls.

The development environment supports several lightweight analysis techniques, such as type checking (Contribution 1) and model visualization support (Contribution 2). Other validation strategies, such as variable counting and connectedness analysis, were implemented (in collaboration with Gábor Bergmann and Ádám Dudás, a BSc student under our joint supervision) to find other typical issues with query definitions.

Tools II: VIATRA Model Transformation Framework

The VIATRA model transformation framework is an open source Eclipse project featuring a complete transformation development environment based on a custom metamodeling core, including a dedicated transformation language and support for both local search-based and incremental query evaluation strategies. I have been serving as co-leader of the project for three years by the the of writing this thesis.

Contribution 1 enabled to extend the VIATRA2 model transformation framework with a type checker implementation, while the results of Contribution 3 were used to extend the debugging capabilities of the framework.

The new version of VIATRA3 provides a high-level model for developing both batch and incremental transformations, and relies on the EMF-IncQuery project for query evaluation, making the project capable of working with EMF-based models natively.

The modular, plug-in based architecture of EMF-IncQuery and VIATRA enables easy integration with existing EMF-based modeling tools and applications, such as Papyrus UML [Papyrus], Capella [Capella], Sirius [Sirius] or Artop [Artop]. Furthermore, while we illustrate EMF-IncQuery and VIATRA in the context of EMF models, core queries and transformations are regular Java programs which have been successfully adapted to other technological spaces (outside Eclipse), such as for Matlab [Hor+14; HRS15] or MPS [TV16].

The projects have already been successfully used in various industrial projects, including, but not limited to the development of an incremental code generator for executable UML models [Hor+15] and well-formedness validation of AutoSAR models in the automotive domain [Ber+10]. These projects both motivated the research presented in this dissertation, and benefited from the integrated static analysis or graphical view generation capabilities presented here.

Applications I: Incremental Graphical Views for Integrated Modular Avionics

The results of Contribution 2 are applied in Trans-IMA, a cooperative research project between Embraer (the large Brazilian airframer) and our research group. The aim of the Trans-IMA project [Hor+14] was to develop a modeling tool that allows engineers to automatically generate integrated Matlab Simulink models for analyzing the complete hardware-software architecture of the system using a high-level allocation process mapping functions to their computing units of the hardware platforms.

The proposed approach uses a set of models describing the required avionics functions (Functional Architecture Model, FAM) and the underlying hardware architecture (Platform Description
Model, PDM) as inputs that are combined in an Allocation Specification Editor requiring inputs from
the developer, then an allocation process generates an Integrated Architecture Model (IAM) that
stores all Simulink-specific configurations, such as library links and model references. All these
models are available in a set of graphical editors with multiple viewpoints that highlight
different segments of the underlying model.

These viewpoint visualizations were built using the Graphical Viewers functionality (Con-
tribution 2), relying on the yFiles for Java \cite{WEK04} graph visualization library, supporting (1)
incremental layout algorithms scaling to thousands of nodes and edges, (2) hierarchical graphs
and closable group nodes and (3) detailed fine-tuning options of layout and visualization param-
eters.

The declarative definition of the visualization supported quick prototyping (using the default
display options provided by the graph viewers). After the set of promising graphical views were
collected, the visualizations could be customized using the Java API of yFiles, such as selecting
colors, layout algorithms, etc.

Applications II: Using Model Queries for Software Maintenance

In a cooperative project with Refactoring 2011 Kft. the applicability of model query and transfor-
mation tools was evaluated for software code refactoring. The project aimed to reduce the effects
of software erosion by detecting coding anti-patterns and propose automatic refactorings that
fix these issues, but leaving it up to the developer to decide how to handle the revealed issues.
The process relied on the capabilities of the Columbus reverse engineering tool \cite{Fer+02}.

The proposed refactoring workflow consisted of four steps: (1) creating an abstract syntax
graph (ASG) from the source code using Columbus, (2) searching for predefined anti-patterns in
this syntax graph, (3) executing the wanted refactorings on the ASG, and finally (4) writing back
the changed model into the original source code. As steps (2) and (3) could benefit from high-
level specification languages used for model query and transformations they were evaluated by
comparing a traditional, Java-based implementation with another relying on EMF-IncQuery and
ViATRA.

The benchmarking has shown that the different approaches complement each other well: a
manually optimized model representation requires less memory, allowing to load larger models
into the memory, while indexing-based techniques, such as the Rete algorithm in EMF-IncQuery
perform well if the anti-patterns have to be re-evaluated after incremental model updates.

For the identification of the smells a selection of graph pattern were implemented in EMF-
IncQuery, where the enhancements to the development environment and the graphical viewers
support for debugging eased development. Furthermore, the created performance benchmark
allows to measure the scalability of different pattern matching strategies on real-world data sets.

The conference paper describing the benchmarking \cite{9} received the best paper award at the
IEEE CSMR-WCRE 2014 Software Evolution Week, Antwerp, Belgium, February 3-6, 2014, and
an extended version was published in the Journal of Information and Software Technology \cite{2}.

5 Publication list of Zoltán Ujhelyi

Number of publications: 26
Number of peer-reviewed publications: 20
Number of independent citations: 43
International, peer-reviewed journal papers


International, peer-reviewed conferences


Zoltán Ujhelyi


International, peer-reviewed workshops


National conferences


[22] Z. Ujhelyi. “Static Type Checking of Model Transformation Programs”. In: Conference of PhD Students in Computer Science. 2010, pp. 413–415


Technical reports and online content


Acknowledgements

My research work was partially supported by the EU projects CONCERTO (ART-2012-333053), DIANA (AERO1-030985), MONDO (ICT-611125) and SECURECHANGE (ICT-FET-231101), the Hungarian CERTIMOT (ERC_HU-09-1-2010-0003) project, the TÁMOP (4.2.2.B-10/1–2010-0009) grant, the GOP (1.2.1-11-2011-0002) grant and a collaborative project with Embraer.
References


Zoltán Ujhelyi PhD thesis booklet


