Project in software engineering ECCO KeFaX (ECCO Linux KErnel FeAture eXtraction) - documentation

Harald A. Weiner
JKU
Linz, Austria
Email: harald.weiner@jku.at

Abstract

This project aims to use a C code parser to walk through the produced abstract syntax tree (AST) to provide an importer to the *ECCO* (*Extraction and Composition for Clone-and-Own*) tool. The goal is to offer case studies for the feature model exploration implementations in *ECCO*. Various approaches have been explored and finally Eclipse MoDisco in combination with the Xtext plug-in have been chosen to reverse-engineer C programs and import them into the EMF (Eclipse Modeling Framework).

I. Introduction

This project has been developed as part of my project in software engineering for the master course computer science. The tool *ECCO* (*Extraction and Composition for Clone-and-Own*)[1] is developed at the ISSE (Institute for Software Systems Engineering [2]) at Johannes Kepler University in Linz, Austria and can be found at [3]. It maps commonalities and differences of existing variants of a portfolio to a feature set. To provide a real-life case study, the Linux Kernel has been chosen as a demonstration example for the *ECCO* tool (the Linux kernel is a pretty well-known example / case study and a project with huge impacts in industry and research). At the moment, plain C and C++ programs are not supported by *ECCO* yet. Therefore, an importer should be written which is able to use the output of C parsers to extract the relevant information for *ECCO*. The KeFaX project should provide an importer for the Linux kernel to the ECCO tool. As a result, this project is supposed to parse the Linux .config file, set-up the minimal infrastructure of source code for the modules and parse the source code. At the end, this project should create an "input tree" data structure. Various approaches exist which have been evaluated for their suitability to the given task. This paper will present the steps taken so far. The next chapter will provide a short description of which steps are necessary to use KeFaX or to inspect the source code. Then an en-detailed discussion can be started. The third chapter will provide an overview of the Linux Kernel compilation and building process. The forth and fifth sections will explain the requirements for this project and finally the sixth chapter will document the implementation phase.

May 09, 2016

II. Getting started

This project is provided as a set of Eclipse plug-ins. Depending on if you would like to use the project or just want to explore the source code there are different requirements. Both development and runtime execution have been tested under *Eclipse Modeling Luna SR2* and *Eclipse Modeling Mars.2 Release (4.5.2)* on AMD64 architecture with a Linux operating system and at least 8 GB RAM.

Warning: This is a prototype / proof-of-concept and is not intended to be used in production environments!!! It may contain some serious bugs, security issues or design flaws which might lead to data loss or data corruption. You have been warned ;-).

A. How-To use

1) Ensure that you have Git installed and that the git executable is in your current $PATH variable.
2) Download Eclipse Modeling IDE from
   - either https://eclipse.org/downloads/[4]
   - or https://eclipse.org/downloads/packages/release/luna/sr2[5]
3) Unzip and open the Eclipse IDE
4) Install Eclipse Modisco (Help→Install new software, select the predefined software site *Modeling package updates for Eclipse Mars*[6] or *Modeling package updates for Eclipse Luna*[7] and install either *Modisco/Modisco SDK (incubation) 0.13.2.201601200708* or *Modeling/Modisco SDK (Incubation) 0.12.2.201501021045* (depending on your Eclipse version).
5) Install NeoEMF by opening the install new software dialog again and add https://timeraider4u.github.io/NeoEMF/[8] as NeoEMF update site. Install

- *Base/NeoEMF Persistence framework*
- *Backends/NeoEMF Blueprints adapter*
- and *Backends/NeoEMF Blueprints implementation*

each with version 0.0.1.2016040202

6) Install *org.xtext.antlr.generator* by adding https://timeraider4u.github.io/org.xtext.antlr.generator/[9] as an update site and selecting the feature *org.xtext.antlr.generator/org.xtext.antlr.generator.feature* with version 3.2.1.201604141818.
7) Install the modified version of *Xtext* by adding https://timeraider4u.github.io/xtext/[10] as an update site and select *Xtext/Xtext Complete SDK 2.9.0.v201604150031*
8) Install KeFax by adding https://timeraider4u.github.io/kefax/[11] as an update site and select *at.jku.weiner.kefax/at.jku.weiner.kefax* with version 0.1.0.201605080110
9) Your installed software (Help→About Eclipse→Installation Details) should now look similiar to the screenshot shown in figure 1:

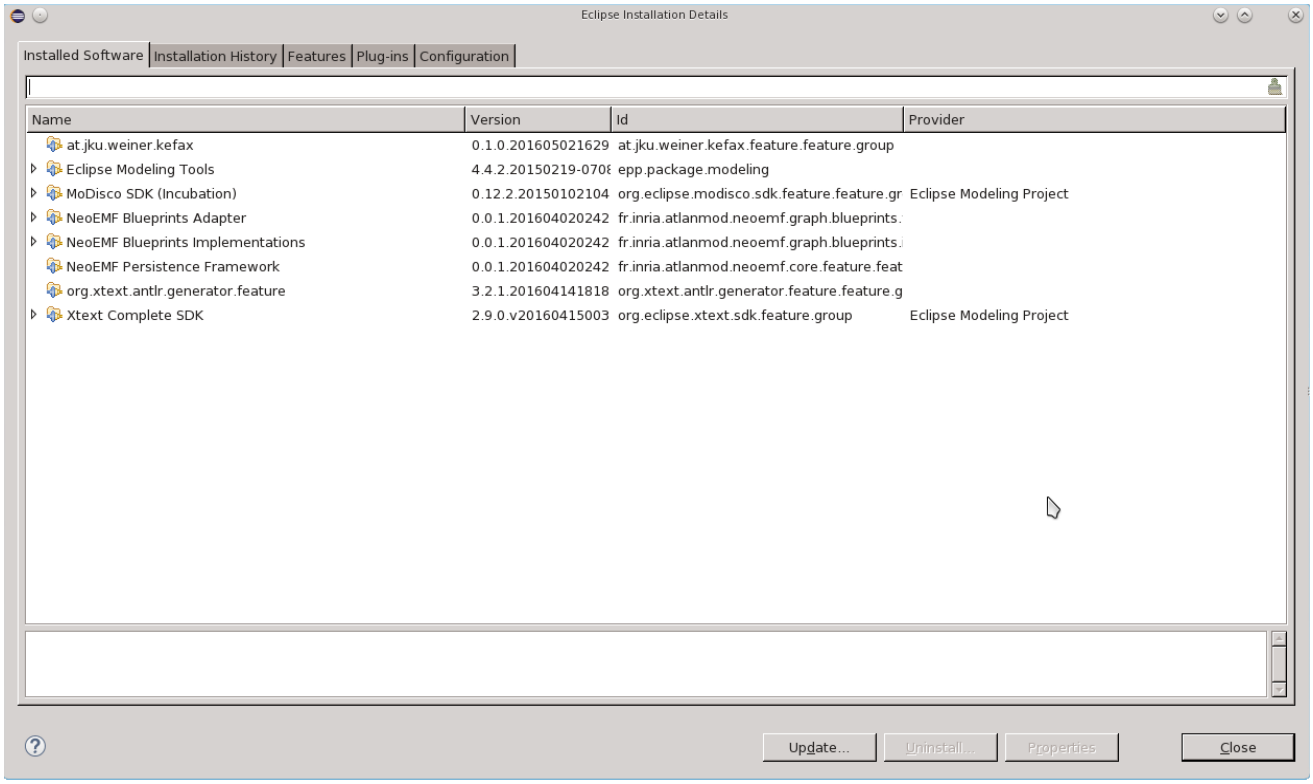

Fig. 1. Installation details

10) Edit the eclipse.ini file. It should contain the following configuration:

Listing 1. part of the eclipse.ini

```
--launcher.XXMaxPermSize
512m
--launcher.defaultAction
openFile
--launcher.appendVmargs
-vmargs
-Dosgi.requiredJavaVersion=1.7
-XX:MaxPermSize=512m
-Xms512m
-Xmx2800m
```

11) Restart Eclipse
12) Run by selecting menu items from *KeFaX* menu (shown in figure 2) either

- KeFax →Run KeFax demonstration A
- or KeFax →Run KeFax demonstration B

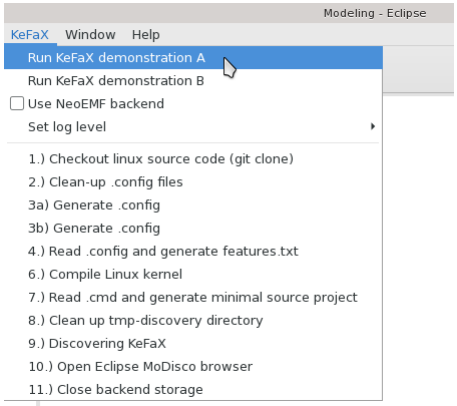

Fig. 2. Using KeFaX after installation

This will now download the Linux source code with git, generating a minimal working configuration file, execute a kernel compilation to obtain the compile options for all source files, generate a *features.txt* file in the destination project *kefax-linux-working*, copy the minimal required source and header files to the *kefax-linux-working project* and start discovering the *kefax-linux-working project*. Once pre-processing and parsing is done, KeFax will open the *MoDisco EMF browser* which shows the reverse-engineered Linux kernel C source model.

Demonstration mode A and demonstration mode B just differ by one feature: B has *CONFIG_UNIX98_PTYS* set to yes will while is not set for A at all. This is either done in step *3a) Generate .config* or in step *3b) Generate .config* of the *KeFaX* menu.

The resulting EMF model file(s) can than be found in the folder *tmp-discover* of the *kefax-linux-working* project.

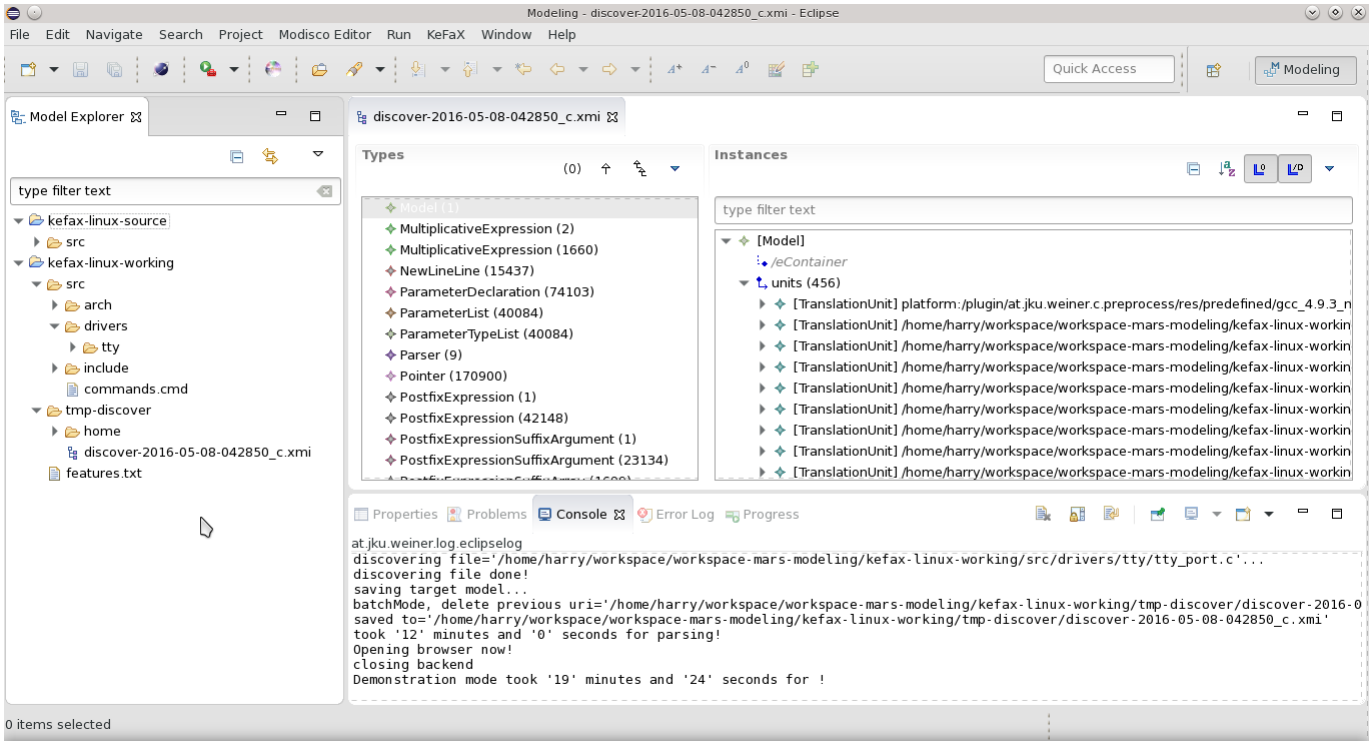

Fig. 3. Result after running KeFaX

You might also adjust the log-level or run the individual commands step-wise to see what they are doing in detail.

Warning: Do not run with log-level set to *trace*. Log-level *trace* is only meant to be used for debugging very nasty bugs (e.g., preprocessor macro expansion). It will take almost forever to execute the preprocessor due to printing the detailed log to the console. Use at your own risk... You have been warned ;-).

B. How-To develop

The whole project is distributed under Eclipse Public License - v 1.0 [12] unless otherwise stated. The source code can be obtained with git from https://github.com/timeraider4u/kefax[13]

1) `git clone https://github.com/timeraider4u/kefax`
2) Execute steps 2 – 7 from How-To use II-A
3) Add https://timeraider4u.github.io/kefax/[11] as an update site, just like in step 8 of How-To use, but instead of installing *kefax* select *at.jku.weiner.xtexttest/at.jku.weiner.xtexttest version 0.1.0.201605080110*
4) Restart Eclipse
5) Import project into workspace (File → Import → General → Existing projects into workspace) and select the workspace folder inside the local *kefax* git repository as the root directory. Select all projects and start the import process.
6) 6. If there are any errors/failures shown after importing you may try to execute Project → Clean → Clean all projects. This will remove temporary Xtext/Xtend files and enforce a global rebuild.
7) The code structuring will be explained later in this paper.
8) KeFaX uses Maven (and Github Travis) for continuous integration: A local Maven 3.0 build can be started by navigating to the local kefax git repository root and executing
9) `mvn clean install`

   . This will also execute all JUnit tests.
10) Feel free to start a pull request or report an issue on the Github page [13].
    - The *master* branch is used for development
    - The *gh-pages* branch is used to store the Eclipse update site.
11) Also take a look at the *README.md* file and execute git pull from time to time to keep in touch with the latest changes.

III. Linux kernel configuration/build

The Linux kernel is developed by programmers from all around the world and can be obtained at [14]. A mirrored version of the Linux kernel can be found on Github [15]. The kernel is developed, compiled and installed by using Unix-like tools. The main documentation is provided as plain text files while the kernel itself is programmed in the C programming language. Although the kernel itself is developed in C and Assembler, some of the included helper tools are written in C++, in Bash shell scripts, or even in Python (e.g., see *tools/perf/python/*) or in Perl (e.g., see *tools/perf/perl*). The Linux kernel also uses its own ecosystem of "programming languages" for configuration and building of the binary objects which are somehow "domain-specific languages", which have historically grown over time.

The picture 4 shows an overview of the dependencies between the different "DSLs". An overview of the kernel compilation and building can also be found at [16]

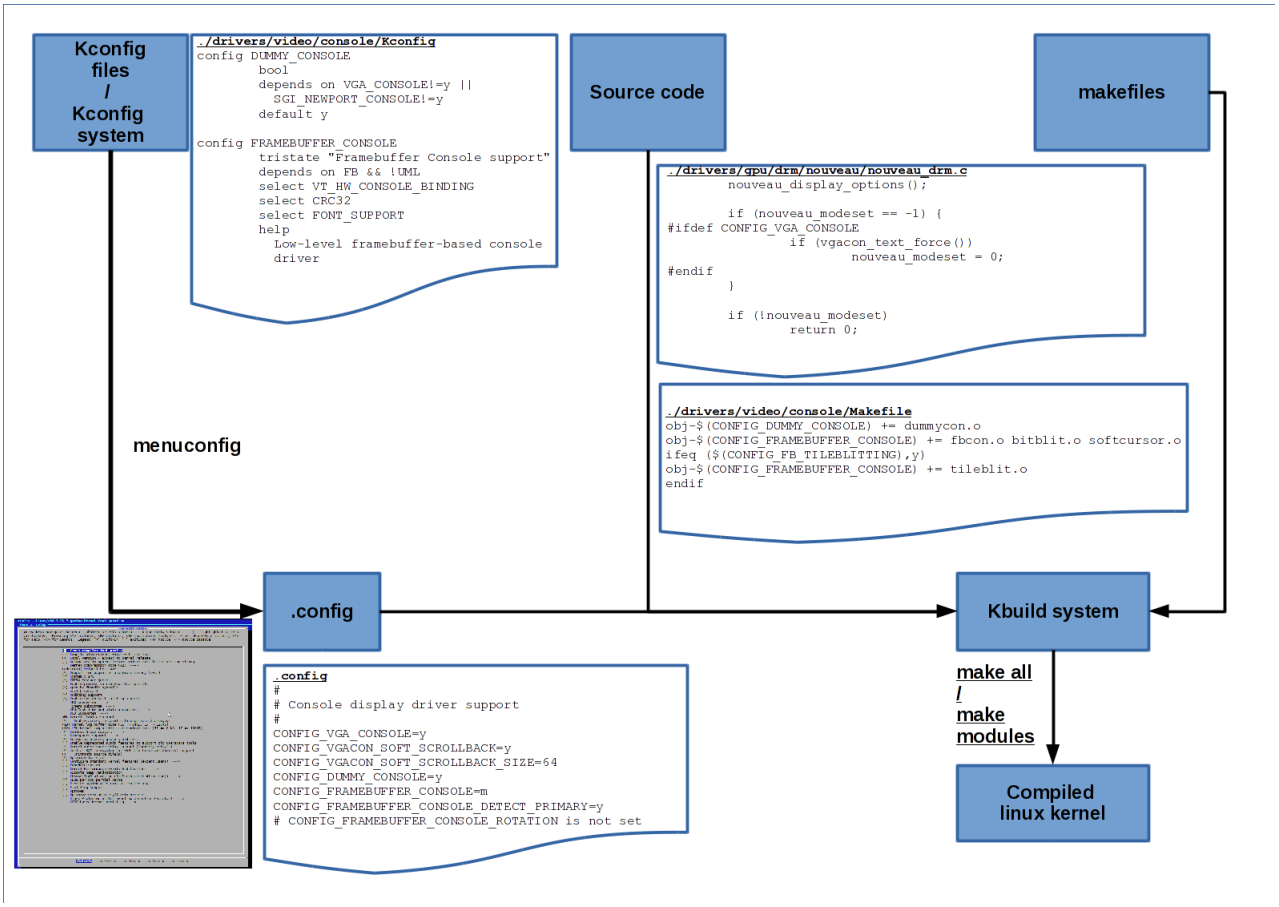

Fig. 4. Overview over the Linux kernel configuration and build process

A. *.config*

First, the Kconfig system specifies which features depend on each other or can not be combined together. Therefore, kconfig serves as some kind of non-formalized feature model. Here is an example:

Listing 2. ./drivers/video/console/Kconfig

```
config DUMMY_CONSOLE
        bool
        depends on VGA_CONSOLE!=y || SGI_NEWPORT_CONSOLE!=y
        default y

config FRAMEBUFFER_CONSOLE
        tristate "Framebuffer Console support"
        depends on FB && !UML
        select VT_HW_CONSOLE_BINDING
        select CRC32
        select FONT_SUPPORT
        help
          Low-level framebuffer-based console driver
```

A detailed description of the kconfig system can be found at [17]

The variability model of the Linux kernel is, e.g., described in She et al. “The Variability Model of The Linux Kernel”[18], in Sincero et al. “The Linux Kernel Configurator as a Feature Modeling Tool”[19] and in Tartler et al. “Dead or alive: finding zombie features in the linux kernel ”[20]. Although these papers describe older kernel versions their main conclusions are still valid.

B. Generating a .config file

The users can then use one of the configuration tools which get delivered with the Linux kernel, e.g. *make oldconfig* when you already have a running kernel and you want to only update to a new kernel version or *make menuconfig*, etc. These applications use the kconfig files as input to determine which additional features must be selected or hide options which are not available due to conflicts. The configuration is then saved into a *.config* file at the root directory of the Linux kernel source code.

The following text listing is an example for such a generated configuration file:

Listing 3. .config file snippet

```
#
# Console display driver support
#
CONFIG_VGA_CONSOLE=y
CONFIG_VGACON_SOFT_SCROLLBACK=y
CONFIG_VGACON_SOFT_SCROLLBACK_SIZE=64
CONFIG_DUMMY_CONSOLE=y
CONFIG_FRAMEBUFFER_CONSOLE=m
CONFIG_FRAMEBUFFER_CONSOLE_DETECT_PRIMARY=y
# CONFIG_FRAMEBUFFER_CONSOLE_ROTATION is not set
```

The following screenshot in figure 5 shows the *menuconfig* program, a graphical configuration tool which can be used on the terminal:

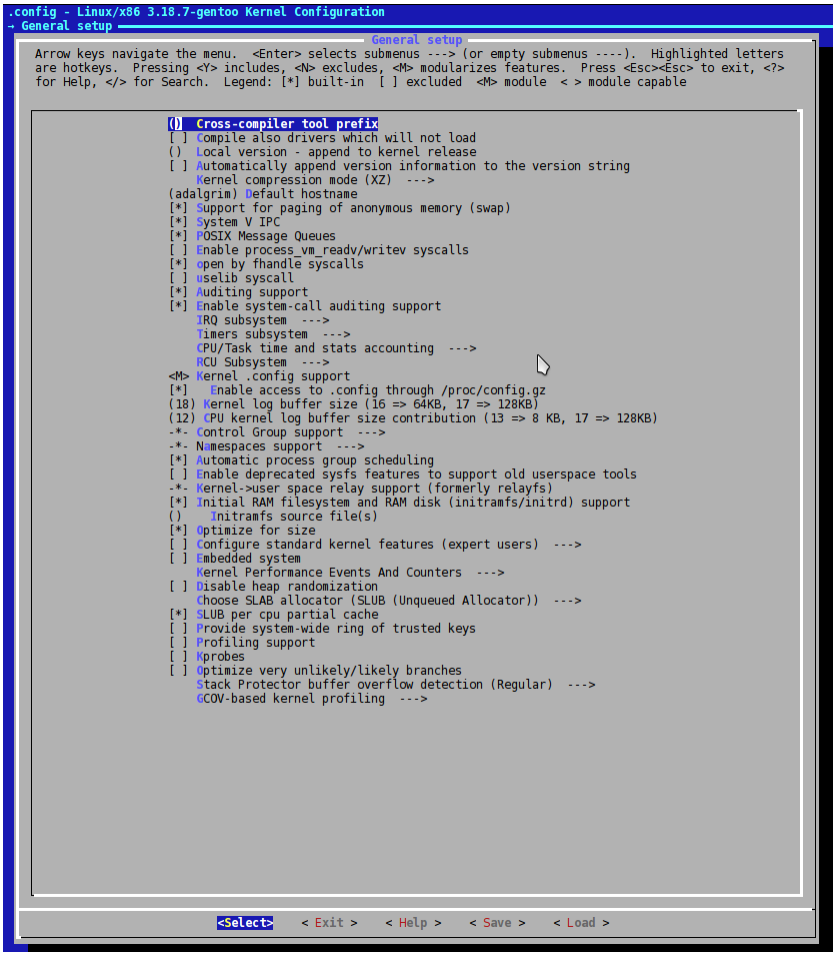

Fig. 5. Configuring the kernel with menuconfig, e.g., over a ssh connection

C. Kbuild system

Whenever the kernel, its modules or any initramfs should be built the kbuild system is invoked. Some valid targets for building are

- make all
- make kernel
- make modules
- make vmlinux
- make initramfs

The makefile in the root directory of the linux source code then invokes the kbuild which generates / collects all necessary makefiles and processes them step-wise. Most of them just contain plain GNU make instructions but some contain additional bash script magic.

An example kbuild makefile is the following:

Listing 4. Snippet from ./drivers/video/console/Makefile

```
obj-$(CONFIG_DUMMY_CONSOLE) += dummycon.o
obj-$(CONFIG_FRAMEBUFFER_CONSOLE) += fbcon.o bitblit.o softcursor.o
ifeq ($(CONFIG_FB_TILEBLITTING),y)
obj-$(CONFIG_FRAMEBUFFER_CONSOLE) += tileblit.o
endif
```

The commands in the makefile then instruct the C compiler which source files to compile and which command line options to use. The selected features of the *.config* file are provided as pre-processor macro definitions. But also include directories are enlisted.

The following source snippet shows how the macro definitions are used in the C source code:

Listing 5. Snippet from ./drivers/gpu/drm/nouveau/nouveau_drm.c

```
        nouveau_display_options();
        if (nouveau_modeset == -1) {
#ifdef CONFIG_VGA_CONSOLE
                if (vgacon_text_force())
                        nouveau_modeset = 0;
#endif
        }
        if (!nouveau_modeset)
                return 0;
```

The results of the compilation are binary object files which can be installed to */boot* and/or */lib/modules/<linux-version>* with *make install*.

More information about kbuild itself can be found at [21] and [22]

IV. Required code structure for ECCO

Ecco requires a structure similiar to the file tree shown in figure 6 to be able to parse the product

The *feature.txt* should list all enabled features, one per line starting with a unique name separated with a semicolon from the feature's description and a new-line character at the end. The file structure should only contain the pruned source code (so no folders / source files for disabled features). The source files should be parsed by some parser and be presented to *ECCO* as a tree data structure.

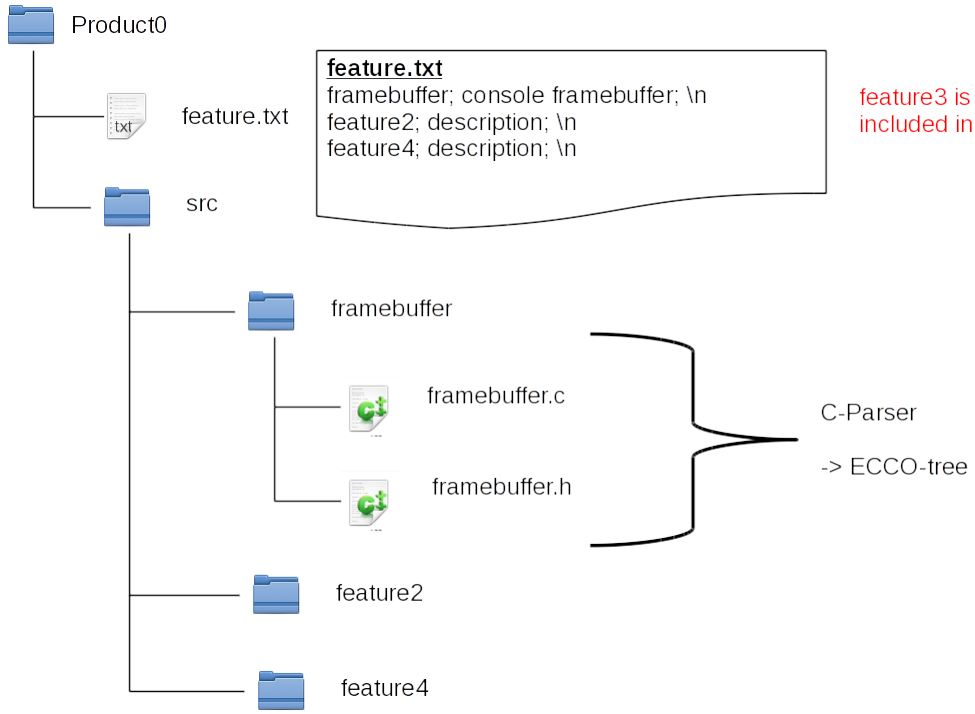

Fig. 6. Overview of required code structure

V. KeFaX requirements / KeFaX tasks

As a result of the required input structure of *ECCO*, the following coarse tasks can be identified for this project:

1) Parsing .config file
2) Obtaining compilation options (which source files are required to be parsed, which include directories are used)
3) Pre-processing source code and removing un-taken pre-processor conditionals
4) Parsing source code (.c / .h files)
5) "Glue code" + (Complete data model)
6) Providing the data structure somehow to *ECCO*

These steps can then be re-defined as:

1) Manually create *.config* by selecting features (e.g., in *menuconfig*)
2) Parse *.config*
3) Create *features.txt* (*.config* $\rightarrow$ *features.txt*)
4) Obtain compilation options from kbuild makefiles
5) Prune source code files (code files for features that are not selected should be deleted) (*.config* + makefiles $\rightarrow$ code structure)
6) Run pre-processor on files and do fine-grain (*# ifdef* code which is not selected should be thrown away) (*.config* $\rightarrow$ source code files)
7) Parse source code and create/import abstract syntax tree.

VI. Implementation

This chapter will clarify the development history, discovered bugs and design decisions taken. Let us first take a look on programs which already existed when this project started and why they have or have not been used for KeFaX and then go into the implementation details.

A. Using existing compilers / compiler-generators

C/C++ are very complex programming languages with a lot of disambiguties, lots of different revisions (e.g. ANSI C, ISO/IEC C99[23], C++11) and a huge amount of compiler specific extensions (GNU GCC extensions, etc.). The Linux kernel

makes use of some of the GNU C extensions. Therefore, it is not possible to compile the vanilla sources of the Linux kernel with Clang LLVM yet - the Clang front-end[24] would provide a nice abstract syntax tree (AST), e.g., as a XML file. Clang claims to be fully GNU gcc compliant and is used as a replacement of the GNU compiler suite already in several projects because of its advanced static code analyzing capabilities. Part of the function declaration of the main- function for the hello-world example can be found in code listing 6.

Listing 6. part of the clang-ast output file

```
-FunctionDecl 0x234e1d0 <helloworld.c:6:1, line:9:1> line:6:5 main 'int (int, char
    **)'
  |-ParmVarDecl 0x234e090 <col:10, col:14> col:14 argc 'int'
  |-ParmVarDecl 0x234e100 <col:20, col:28> col:28 argv 'char **'
  `-CompoundStmt 0x234e400 <col:34, line:9:1>
    |-CallExpr 0x234e360 <line:7:2, col:25> 'int'
    | |-ImplicitCastExpr 0x234e348 <col:2> 'int (*)(const char *, ...)' <
    |   FunctionToPointerDecay>
    | | `-DeclRefExpr 0x234e280 <col:2> 'int (const char *, ...)' Function 0x22fa740
    |   'printf' 'int (const char *, ...)'
    | `-ImplicitCastExpr 0x234e3a8 <col:9> 'const char *' <BitCast>
    |   `-ImplicitCastExpr 0x234e390 <col:9> 'char *' <ArrayToPointerDecay>
    |     `-StringLiteral 0x234e2e8 <col:9> 'char [14]' lvalue "Hello World!\n"
```

Another issue is still that their is only limited support for traversing, working and exporting the AST for further investigations by other applications.

GNU gcc itself provides its internal data structures to external programs. Applications might use this information and provide AST as XML, e.g. XOgastan[25] or GCC_XML [26]. An example output for gcc-xml can be seen in source code listing 7

Listing 7. part of the gccxmloutput.xml file

```
<Function id="_187" name="qfcvt" returns="_368" throw="" context="_1" location="
    f4:835" file="f4" line="835" extern="1" attributes="__nonnull__(,)">
  <Argument name="__value" type="_586" location="f4:835" file="f4" line="835"/>
  <Argument name="__ndigit" type="_69" location="f4:835" file="f4" line="835"/>
  <Argument name="__decpt" type="_694r" location="f4:835" file="f4" line="835"/>
  <Argument name="__sign" type="_694r" location="f4:835" file="f4" line="835"/>
</Function>
<Function id="_188" name="mbtowc" returns="_69" throw="" context="_1" location="
    f4:867" file="f4" line="867" extern="1">
```

When at first this method seemed promising, it turned out that gcc-xml is generating too much non-descriptive IDs, is sometimes buggy and that the GNU gcc itself is not producing an easily iterate-able AST itself.

Compiler compilers are based on processing formal grammars. The research in theoretic computer science has led to the automation of generating scanners and parsers out of an existing formal description text of the programming language (wherefore often the EBNF - extended Backus-Naur-Form - is used as a notation). For compiler generators like Coco/R[27] (which has been developed at the SSW institute at the JKU in Linz), GNU's implementation of Yacc called Bison[28], etc. there is no unique and/or clear EBNF available for C/C++ which would also include most of the GNU C extensions. One disadvantage of compiler compilers is the mixture of multiple languages (one language for describing the scanning/parsing process mixed with the source code statements for the resulting compiler). Another disadvantage is that often C/C++ EBNF descriptions lack language features like the GNU C extensions or others. But the worst matter is that all information about preprocessor statements in C and C++ are lost because EBNFs can not deal with include or define macros. EBNF profiles of various programming languages can be found in the grammarware Github repository[29]. A little out-dated grammar description for GNU GCC can be found at [30].

The problem with most of the evaluated tools is that they are either not available anymore or are out-dated or buggy. But even if they would work, the next question is how to build-up the AST, which tools and which data structures to use and how much effort would be required to do so ...

B. Reverse Engineering

How about using *EMF (Eclipse Modeling Framework)* [31] [32] instead? EMF provides rich APIs for processing and transforming a model into other models or code. An tutorial on EMF can, for example, be found at [33]. So which projects

exist for working with C/C++ in conjunction with EMF?

1) EMF4CPP: The first hit on Google when searching for EMF and C/C++ is *EMF4CPP* [34] [35]. It provides the ability to work with EMF models inside a C++ project (just like Java is supported out-of-the-box) by including the library which is part of EMF4CPP. Another feature of this open-source project is to enable the generation of C++ source text out of an Ecore model. It also comes with a XText artifact for importing C++ code into an EMF model. Unfortunately, as the goal of EMF4CPP is a different one, it only supports a small subset of C++, e.g. it is not possible to use macro directives, embedded C or assembler code. EMF4CPP is, therefore, more similar to the *Javascript for EMF (JS4EMF)* project [36] [37]. It is not possible to parse the Linux kernel with this project.

2) Eclipse MoDisco: So far, the approaches have been disappointing because every one had major disadvantages. But the method described in this section about Eclipse MoDisco seemed promising.

The paper "Model-Based Mining of Source Code Repositories" [38] written by Markus Scheidgen and Joachim Fischer describes how they utilized *Eclipse MoDisco* for reverse-engineering of open source code repositories like GitHub to gain static code analysis metrics about Java applications.

Eclipse MoDisco [39] [40] [41]. claims to provide "an extensible framework to develop model-driven tools to support use-cases of existing software modernization".

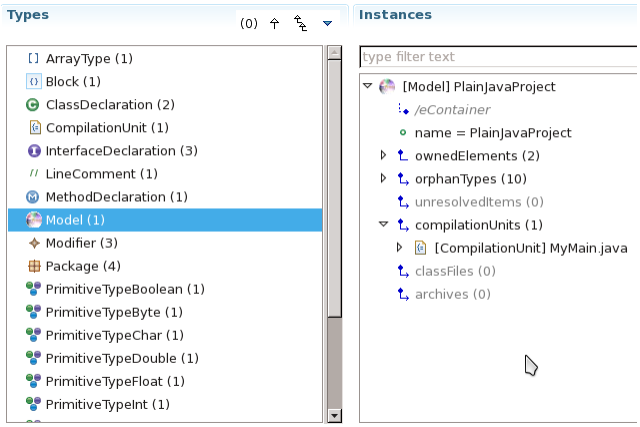

Fig. 7. Eclipse MoDisco model for a simple Java project

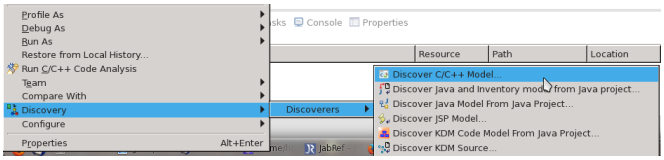

Fig. 8. Using Eclipse MoDisco

It is a reverse-engineering framework which is built on top of the *Eclipse EMF* (Eclipse Modeling Framework). This would solve the question of which data structure / abstract syntax tree to use for storage. Discoverers are a set of Eclipse plug-ins that provide an importer/parser for a certain programming language. There exist discoverers for Java, JSP, JSON and many more. Unfortunately, *Eclipse MoDisco discoverers* for C and C++ did not exist when starting this project. In the Eclipse forum in a thread called "Looking for c/c++ discoverer" [42] it was suggested to piggy-back use *Eclipse CDT ((Compiler development tools))* [43][44] to implement a C/C++ discoverer.

Eclipse CDT supports C/C++ and can deal with most of the GNU C extensions and pre-processing directives but it has a deeply nested code structure that is not always easy to deal with when walking the tree, as it can be seen in code snippet 8.

Listing 8. Javadoc subinterfaces of IASTNode

```
All Known Subinterfaces:
    IASTArrayDeclarator, IASTArrayModifier, IASTArraySubscriptExpression,
        IASTASMDeclaration, IASTBinaryExpression,
    IASTBinaryTypeIdExpression, IASTBreakStatement, IASTCaseStatement,
        IASTCastExpression, IASTComment,
```

IASTCompositeTypeSpecifier , IASTCompoundStatement , IASTConditionalExpression ,
IASTContinueStatement , IASTDeclaration ,
IASTDeclarationListOwner , IASTDeclarationStatement , IASTDeclarator ,
IASTDeclSpecifier , IASTDefaultStatement , IASTDoStatement ,
IASTElaboratedTypeSpecifier , IASTEnumerationSpecifier , IASTEnumerationSpecifier.
IASTEnumerator , IASTEqualsInitializer ,
IASTExpression , IASTExpressionList , IASTExpressionStatement , IASTFieldDeclarator ,
IASTFieldReference , IASTForStatement ,
IASTFunctionCallExpression , IASTFunctionDeclarator , IASTFunctionDefinition ,
IASTFunctionStyleMacroParameter , IASTGotoStatement ,
IASTIdExpression , IASTIfStatement , IASTImplicitName , IASTImplicitNameOwner ,
IASTInitializer , IASTInitializerClause ,
IASTInitializerExpression , IASTInitializerList , IASTLabelStatement ,
IASTLiteralExpression , IASTName, IASTNamedTypeSpecifier ,
IASTNullStatement , IASTParameterDeclaration , IASTPointer , IASTPointerOperator ,
IASTPreprocessorElifStatement ,
IASTPreprocessorElseStatement , IASTPreprocessorEndifStatement ,
IASTPreprocessorErrorStatement ,
IASTPreprocessorFunctionStyleMacroDefinition , IASTPreprocessorIfdefStatement ,
IASTPreprocessorIfndefStatement ,
IASTPreprocessorIfStatement , IASTPreprocessorIncludeStatement ,
IASTPreprocessorMacroDefinition , IASTPreprocessorMacroExpansion ,
IASTPreprocessorObjectStyleMacroDefinition , IASTPreprocessorPragmaStatement ,
IASTPreprocessorStatement ,
IASTPreprocessorUndefStatement , IASTProblem , IASTProblemDeclaration ,
IASTProblemExpression , IASTProblemStatement ,
IASTProblemTypeId , IASTReturnStatement , IASTSimpleDeclaration ,
IASTSimpleDeclSpecifier , IASTStandardFunctionDeclarator ,
IASTStatement , IASTSwitchStatement , IASTTranslationUnit , IASTTypeId ,
IASTTypeIdExpression , IASTTypeIdInitializerExpression ,
IASTUnaryExpression , IASTWhileStatement , ICASTArrayDesignator , ICASTArrayModifier
, ICASTCompositeTypeSpecifier ,
ICASTDeclSpecifier , ICASTDesignatedInitializer , ICASTDesignator ,
ICASTElaboratedTypeSpecifier , ICASTEnumerationSpecifier , I
CASTFieldDesignator , ICASTKnRFunctionDeclarator , ICASTPointer ,
ICASTSimpleDeclSpecifier , ICASTTypedefNameSpecifier ,
ICASTTypeIdInitializerExpression , ICPPASTAmbiguousTemplateArgument ,
ICPPASTArrayDeclarator , ICPPASTArraySubscriptExpression ,
ICPPASTBinaryExpression , ICPPASTCapture , ICPPASTCastExpression ,
ICPPASTCatchHandler , ICPPASTCompositeTypeSpecifier ,
ICPPASTCompositeTypeSpecifier.ICPPASTBaseSpecifier ,
ICPPASTConstructorChainInitializer , ICPPASTConstructorInitializer ,
ICPPASTConversionName , ICPPASTDeclarator , ICPPASTDeclSpecifier ,
ICPPASTDeleteExpression , ICPPASTElaboratedTypeSpecifier ,
ICPPASTEnumerationSpecifier , ICPPASTExplicitTemplateInstantiation ,
ICPPASTExpressionList , ICPPASTFieldDeclarator ,
ICPPASTFieldReference , ICPPASTForStatement , ICPPASTFunctionCallExpression ,
ICPPASTFunctionDeclarator , ICPPASTFunctionDefinition ,
ICPPASTFunctionTryBlockDeclarator , ICPPASTFunctionWithTryBlock ,
ICPPASTIfStatement , ICPPASTInitializerList , ICPPASTLambdaExpression ,
ICPPASTLinkageSpecification , ICPPASTLiteralExpression , ICPPASTNamedTypeSpecifier ,
ICPPASTNamespaceAlias , ICPPASTNamespaceDefinition ,
ICPPASTNewExpression , ICPPASTOperatorName , ICPPASTPackExpansionExpression ,
ICPPASTParameterDeclaration , ICPPASTPointerToMember ,
ICPPASTQualifiedName , ICPPASTRangeBasedForStatement , ICPPASTReferenceOperator ,
ICPPASTSimpleDeclSpecifier ,

```
ICPPASTSimpleTypeConstructorExpression , ICPPASTSimpleTypeTemplateParameter ,
    ICPPASTStaticAssertDeclaration ,
ICPPASTSwitchStatement , ICPPASTTemplateDeclaration ,
    ICPPASTTemplatedTypeTemplateParameter , ICPPASTTemplateId ,
ICPPASTTemplateParameter , ICPPASTTemplateSpecialization , ICPPASTTranslationUnit ,
    ICPPASTTryBlockStatement ,
ICPPASTTypeId , ICPPASTTypeIdExpression , ICPPASTTypenameExpression ,
    ICPPASTUnaryExpression , ICPPASTUsingDeclaration ,
ICPPASTUsingDirective , ICPPASTVisibilityLabel , ICPPASTWhileStatement ,
    IGCCASTArrayRangeDesignator , IGCCASTSimpleDeclSpecifier ,
IGNUASTCompoundStatementExpression , IGNUASTTypeIdExpression ,
    IGNUASTUnaryExpression , IGPPASTBinaryExpression , IGPPASTDeclSpecifier ,
IGPPASTExplicitTemplateInstantiation , IGPPASTPointer , IGPPASTPointerToMember ,
    IGPPASTSimpleDeclSpecifier
```

Therefore, it was decided to change from *Eclipse CDT* to a different parsing technology.

3) ANTLR, EMFText and Xtext: Luckily there is an *ANTLR* grammar available at the *antlr/grammars-v4* Github repository [45]. which already includes most of the GNU C extensions. "ANTLR (ANother Tool for Language Recognition) is a powerful parser generator for reading, processing, executing, or translating structured text or binary files. It's widely used to build languages, tools, and frameworks. From a grammar, ANTLR generates a parser that can build and walk parse trees."[46] There exist bindings for different programming languages including Java (in which most of its parts are also developed) but *ANTLR* does not support exporting to *Eclipse EMF*.

ANTLR itself is used by two Text-To-Model tools for the *Eclipse IDE*: *EMFText* and *XText*. Both provide the possibility to engineer DSLs and use the automated generation of advanced text editors which support syntax highlighting, error recovering and auto-completion. *XText* has the advantage compared to *EMFText* that it is an official Eclipse project and that it can automatically generate an Ecore meta-model out of a grammar description file. Both use a version three of *ANTLR* [47] [48] which has some disadvantages compared to the newer re-written *ANTLR v4* which supports some automatically left-recursion resolution and, therefore, does not need semantic predicates anymore. Even if almost all complex EBNF grammars use left-recursions, most of them can be re-factored to avoid disambiguates.

But there are cases in the C programming languages which require semantic predicates (e.g., to distinguish between declarations and function definitions) to remove left-recursion from the grammar. For more information on semantic predicates see [49] and [50]. *ANTLR* in version four does not need semantic predicates anymore for parsing C/C++, version three supports them at least. Unfortunately, even if *EMFText* and *XText* both use *ANTLR* in version three, they do not support semantic predicates (they only support syntactical predicates).

XText internally generates an ANTLR grammar file (e.g., *InternalParser.g*) out of the DSL grammar description file (e.g., *Parser.xtext*). When semantic predicates where inserted into this intermediate file and this grammar was then fed into the *ANTLR* application as an input, it was possible to create a valid C parser. Therefore, it was necessary to add semantic predicate support to *XText* first, to avoid patching of the intermediate file all the time. This has been implemented in https://github.com/timeraider4u/xtext [51] which is a fork of the official *Xtext* source code. When this was done, the need for an own C pre-processor has arisen which should be able to also gain information on which conditionals where taken, to resolve pre-processor macros and so on... The C pre-processor and parser have then been tested against various test code and parts of the libc implementation. Afterwards the C pre-processor and parser combination have been integrated into an *Eclipse MoDisco* discoverer and the final result has been tested against the Linux kernel. Whenever an error occurred new JUnit test cases have been added and the grammar files have been extended to support these language features as well. Not all GNU C extensions are supported yet (they are far too many).

Unfortunately, some issues still remain unsolved at this moment, including some cases that can lead to *OutOfMemory exceptions*. Even moving from the default *EMF XMI storage back-end* to *NeoEMF* could not resolve this problem satisfactorily. *NeoEMF*[52] (previously called *Neo4EMF*) [53] is a storage back-end for *EMF* models. It provides adapters for *MapDB* and the graph database *Neo4J* to store/deal with *EMF* models.

Some analysis with the Eclipse Memory Analyzer (MAT) [54] [55] has shown that there seem to be two reasons: The first one is that XText uses its own lazy loading implementation which conflicts with using a separate storage back-end (e.g., see [56]). The second reason is that *Eclipse MoDisco* currently only supports *XMI* serialization out-of-the-box and, therefore, the *EMF model browser* can not deal with the lazy loading algorithm (which is used by *NeoEMF*) correctly.

C. MORE TO DO

More detailed descriptions coming soon...

VII. FAQ - FREQUENTLY ASKED QUESTIONS

This section lists some frequently asked questions (or what I suppose to be interesting facts for the public...

Question	Answer
Can I use another (probably newer) version of *org.xtext.antlr.generator* or *Xtext*?	No! There have been several changes made to these software products which are the moment not available in upstream. There are also some bugs which prevent me from efficiently getting these changes d'accord with the master branches, e.g. see bug reports at [57] [58] which themselves required that a bug/feature request in the *EMF genmodel* generator[59] needs to be fixed first. Anyway, even if I could start a pull request for the changes made so far, it would be questionable if they get integrated into the git head at anytime.
Can I use another (probably newer) version of *Eclipse MoDisco* or *NeoEMF*?	Yes, you might give them a try. Please report if newer versions work / do not work for you. You may use the official update site for *NeoEMF* because the issues reported in [60] and [61] have already been merged into the master branch and made their way into the *NeoEMF* official update site.

VIII. LESSONS LEARNED

Asummary of the lessons learned so far ...

- The C programming language has a complex grammar, even without pre-processing directives and compiler extensions, with lots of disambiguties.
- Therefore, there is no *Text-To-Model* tool availabel yet which supports such a complex general programming language like C.
- A tool based on *ANTLR v4* would be great, as *ANTLR v3* no longer receives support and version 4 is a lot more convinient (e.g., semantic predicates no longer needed, tree view on default, AST tree walker moved out of grammar itself, etc.). Also more grammars for version 4 exist and would, therefore, work out-of-the-box.
- It would have been nice to have a dedicated testing DSL for engineering DSLs. When constructing a DSL it is often necessary to check each step of the translation process (lexing, parsing and code generation) in detail. But the test cases get huge and complex when trying to assert that lexer tokens, AST tree structures and the output files are each correct. The DSL used for testing *Xtext* projects used in this project called *XtextTest* has been a workaround for this situation, but it is obviously a poor solution to the problem.
- Another interesting problem for further research would be the co-evolution of abstract and concrete syntax (together with automated refactoring of test cases) during developing a DSL. E.g., you write your grammar and test it on files. Then you have to change the grammar a bit to also include some special cases. As a result of the improvements made to the DSL, the metamodel has to be changed too and then the code generation is not working anymore. After fixing the code generation you find out that also several dozens or even more of your JUnit tests have to be adapted too to reflect the newest features in the latest version of the DSL. This is awkward! Some evolutionary semi-automatical approach would be helpful here, maybe the development of several DSLs would make sense here: The first DSL for describing the concrete syntax could be, e.g., an *ANTLR* grammar file. The second DSL describes the abstract syntax (the meta-model). And then there is a third DSL which actually does the mapping between abstract and concrete syntax. And all of them are updated whenever a single line of these DSLs is changed.

REFERENCES

[1] S. Fischer, L. Linsbauer, R. E. Lopez-Herrejon, and A. Egyed, "Enhancing clone-and-own with systematic reuse for developing software variants," in *Software Maintenance and Evolution (ICSME), 2014 IEEE International Conference on*. IEEE, 2014, pp. 391–400.

[2] "Institute for software systems engineering," http://www.jku.at/isse/content, last visited: 2016-05-09. [Online]. Available: http://www.jku.at/isse/content

[3] "Ecco tool website," http://www.isse.jku.at/tools/ecco, last visited: 2016-05-09. [Online]. Available: http://www.isse.jku.at/tools/ecco

[4] "Eclipse modeling download," https://eclipse.org/downloads/, last visited: 2016-05-09. [Online]. Available: https://eclipse.org/downloads/

[5] "Eclipse modeling download luna," https://eclipse.org/downloads/packages/release/luna/sr2, last visited: 2016-05-09. [Online]. Available: https://eclipse.org/downloads/packages/release/luna/sr2

[6] "Modeling package updates for eclipse mars," http://www.eclipse.org/modeling/amalgam/downloads/package/modeling/mars/, last visited: 2016-05-09. [Online]. Available: http://www.eclipse.org/modeling/amalgam/downloads/package/modeling/mars/

[7] "Modeling package updates for eclipse luna," http://www.eclipse.org/modeling/amalgam/downloads/package/modeling/luna/, last visited: 2016-05-09. [Online]. Available: http://www.eclipse.org/modeling/amalgam/downloads/package/modeling/luna/

[8] "Neoemf update site," https://timeraider4u.github.io/NeoEMF/, last visited: 2016-05-09. [Online]. Available: https://timeraider4u.github.io/NeoEMF/

[9] "org.xtext.antlr.generator update site," https://timeraider4u.github.io/org.xtext.antlr.generator/, last visited: 2016-05-09. [Online]. Available: https://timeraider4u.github.io/org.xtext.antlr.generator/

[10] "Modified xtext update site," https://timeraider4u.github.io/xtext/, last visited: 2016-05-09. [Online]. Available: https://timeraider4u.github.io/xtext/

[11] "Kefax update site," https://timeraider4u.github.io/kefax/, last visited: 2016-05-09. [Online]. Available: https://timeraider4u.github.io/kefax/

[12] "Eclipse public license text," http://www.eclipse.org/legal/epl-v10.html, last visited: 2016-05-09. [Online]. Available: http://www.eclipse.org/legal/epl-v10.html

[13] "Kefax source code," https://github.com/timeraider4u/kefax, last visited: 2016-05-09. [Online]. Available: https://github.com/timeraider4u/kefax

[14] "Linux kernel," https://www.kernel.org/, last visited: 2016-05-09. [Online]. Available: https://www.kernel.org/

[15] "Linux kernel (github mirror)," https://github.com/torvalds/linux, last visited: 2016-05-09. [Online]. Available: https://github.com/torvalds/linux

[16] G. Kroah-Hartman, "The kernel configuration and build process," http://www.linuxjournal.com/article/6568, 2003, last visited: 2016-05-09. [Online]. Available: http://www.linuxjournal.com/article/6568

[17] "Kconfig," https://www.kernel.org/doc/Documentation/kbuild/kconfig-language.txt, last visited: 2016-05-09. [Online]. Available: https://www.kernel.org/doc/Documentation/kbuild/kconfig-language.txt

[18] S. She, R. Lotufo, T. Berger, A. Wasowski, and K. Czarnecki, "The variability model of the linux kernel." *VaMoS*, vol. 10, pp. 45–51, 2010. [Online]. Available: http://gsd.uwaterloo.ca/sites/default/files/camera-vamos-20100107.pdf

[19] J. Sincero and W. Schröder-Preikschat, "The linux kernel configurator as a feature modeling tool." in *SPLC (2)*. Citeseer, 2008, pp. 257–260. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.157.2318&rep=rep1&type=pdf

[20] R. Tartler, J. Sincero, W. Schröder-Preikschat, and D. Lohmann, "Dead or alive: finding zombie features in the linux kernel," in *Proceedings of the First International Workshop on Feature-Oriented Software Development*. ACM, 2009, pp. 81–86. [Online]. Available: http://www.infosun.fim.uni-passau.de/spl/apel/FOSD2009/FOSD2009_Printed_Proceedings.pdf#page=89

[21] "Kbuild makefiles," https://www.kernel.org/doc/Documentation/kbuild/makefiles.txt, last visited: 2016-05-09. [Online]. Available: https://www.kernel.org/doc/Documentation/kbuild/makefiles.txt

[22] H. .p, "Makefile and kconfig," http://hemaprathaban.blogspot.co.at/2013/06/makefile-and-kconfig.html, 2013, last visited: 2016-05-09. [Online]. Available: http://hemaprathaban.blogspot.co.at/2013/06/makefile-and-kconfig.html

[23] *ISO/IEC 9899:1990*, ISO/IEC Std., 2007, last visited: 2016-05-09. [Online]. Available: http://www.open-std.org/JTC1/SC22/WG14/www/docs/n1256.pdf

[24] S. Naroff, "New llvm c front-end," http://llvm.org/devmtg/2007-05/09-Naroff-CFE.pdf, 2007, last visited: 2016-05-09. [Online]. Available: http://llvm.org/devmtg/2007-05/09-Naroff-CFE.pdf

[25] G. Antoniol, M. D. Penta, G. Masone, and U. Villano, "Xogastan: Xml-oriented gcc ast analysis and transformations," in *Source Code Analysis and Manipulation, 2003. Proceedings. Third IEEE International Workshop on*, Sept 2003, pp. 173–182.

[26] "Gcc-xml," http://gccxml.github.io/HTML/Index.html, 2012, last visited: 2016-05-09. [Online]. Available: http://gccxml.github.io/HTML/Index.html

[27] "Coco/r," http://ssw.jku.at/Research/Projects/Coco/, last visited: 2016-05-09. [Online]. Available: http://ssw.jku.at/Research/Projects/Coco/

[28] "Gnu bison website," https://www.gnu.org/software/bison/, last visited: 2016-05-09. [Online]. Available: https://www.gnu.org/software/bison/

[29] "Grammar zoo," https://github.com/grammarware/slps/tree/master/topics/grammars, last visited: 2016-05-09. [Online]. Available: https://github.com/grammarware/slps/tree/master/topics/grammars

[30] C. D. James R. Cordy, Andrew J. Malton, "Txl c basis grammar," http://slebok.github.io/zoo/c/gnu/cordy-malton-dahn/extracted/index.html, 2011, last visited: 2016-05-09. [Online]. Available: http://slebok.github.io/zoo/c/gnu/cordy-malton-dahn/extracted/index.html

[31] "Eclipse modeling framework (emf)," https://eclipse.org/modeling/emf/, 2016, last visited: 2016-05-10. [Online]. Available: https://eclipse.org/modeling/emf/

[32] D. Steinberg, F. Budinsky, E. Merks, and M. Paternostro, *EMF: eclipse modeling framework*. Pearson Education, 2008.

[33] M. Koegel and J. Helming, "What every eclipse developer should know about emf," http://eclipsesource.com/blogs/tutorials/emf-tutorial/, 2016, last visited: 2016-05-10. [Online]. Available: http://eclipsesource.com/blogs/tutorials/emf-tutorial/

[34] "Eclipse modeling framework for c++," https://github.com/catedrasaes-umu/emf4cpp, 2015, last visited: 2016-05-10, previously on https://code.google.com/p/emf4cpp/. [Online]. Available: https://github.com/catedrasaes-umu/emf4cpp

[35] A. Senac, D. Sevilla, and G. Martínez, "Emf4cpp: a c++ ecore implementation," *CABOT, Jordi (Hrsg.)*, pp. 98–106, 2010.

[36] "Proposals - javascript for emf (js4emf)," http://www.eclipse.org/proposals/js4emf/, 2009, last visited: 2016-05-10. [Online]. Available: http://www.eclipse.org/proposals/js4emf/

[37] "Javascript for emf," https://jaxenter.com/javascript-for-emf-100265.html, 2010, last visited: 2016-05-10. [Online]. Available: https://jaxenter.com/javascript-for-emf-100265.html

[38] M. Scheidgen and J. Fischer, "Model-based mining of source code repositories," in *System Analysis and Modeling: Models and Reusability*. Springer, 2014, pp. 239–254. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.657.3002&rep=rep1&type=pdf

[39] "Eclipse modisco home," https://eclipse.org/MoDisco/, last visited: 2016-05-09. [Online]. Available: https://eclipse.org/MoDisco/

[40] H. Bruneliere, J. Cabot, F. Jouault, and F. Madiot, "Modisco: a generic and extensible framework for model driven reverse engineering," in *Proceedings of the IEEE/ACM international conference on Automated software engineering*. ACM, 2010, pp. 173–174. [Online]. Available: https://hal.archives-ouvertes.fr/hal-00534450/document

[41] H. Bruneliere, J. Cabot, G. Dupé, and F. Madiot, "Modisco: A model driven reverse engineering framework," *Information and Software Technology*, vol. 56, no. 8, pp. 1012–1032, 2014. [Online]. Available: https://hal-mines-nantes.archives-ouvertes.fr/docs/00/97/26/32/PDF/MoDisco-JournalPaper-IST_Preliminary.pdf

[42] H. A. W. Jim Foscue, Hugo Bruneliere, "Eclipse community forums - eclipse modisco - c/c++ discoverer(looking for c/c++ discoverer)," https://www.eclipse.org/forums/index.php/t/366540/, 2012-2015, last visited: 2016-05-10. [Online]. Available: https://www.eclipse.org/forums/index.php?t=msg&th=366540&goto=892358&#msg_892358

[43] "Eclipse cdt (c/c++ development tooling) website," https://eclipse.org/cdt/, 2016, last visited: 2016-05-10. [Online]. Available: https://eclipse.org/cdt/

[44] D. Piatov, A. Janes, A. Sillitti, and G. Succi, "Using the eclipse c/c++ development tooling as a robust, fully functional, actively maintained, open source c++ parser." *OSS*, vol. 378, p. 399, 2012. [Online]. Available: https://www.researchgate.net/profile/Alberto_Sillitti/publication/266483397_Using_the_Eclipse_CC_Development_Tooling_as_a_Robust_Fully_Functional_Actively_Maintained_Open_Source_C_Parser/links/550693800cf231de0777fec0.pdf

[45] T. Parr, "antlr/grammars-v4 c.g4," https://github.com/antlr/grammars-v4/blob/master/c/C.g4, 2015. [Online]. Available: https://github.com/antlr/grammars-v4/blob/master/c/C.g4

[46] ——, "Antlr website," http://www.antlr.org/, 2016, last visited: 2016-05-11. [Online]. Available: http://www.antlr.org/

[47] H. A. W. Oleg Bolshakov Sven Efftinge, "Xtext and antlr4?(still no plans to use newer antlr?)," https://www.eclipse.org/forums/index.php/t/640630/, 2016, last visited: 2016-05-11. [Online]. Available: https://www.eclipse.org/forums/index.php/t/640630/

[48] H. A. W. Lars Schtze, Mirko Seifert, "Use antlr 4.0 - issue 17," https://github.com/DevBoost/EMFText/issues/17, 2016, last visited: 2016-05-11. [Online]. Available: https://github.com/DevBoost/EMFText/issues/17

[49] T. J. Parr and R. W. Quong, "Adding semantic and syntactic predicates to ll (k): pred-ll (k)," in *Compiler Construction*. Springer, 1994, pp. 263–277.

[50] M. Juroviov, "Antlr semantic predicates," http://meri-stuff.blogspot.co.at/2012/12/antlr-semantic-predicates.html, 2012, last visited: 2016-05-11. [Online]. Available: http://meri-stuff.blogspot.co.at/2012/12/antlr-semantic-predicates.html

[51] H. A. Weiner, “Xtext modified source code,” https://github.com/timeraider4u/xtext, 2016, last visited: 2016-05-11. [Online]. Available: https://github.com/timeraider4u/xtext
[52] “atlanmod/neoemf source repository,” https://github.com/atlanmod/NeoEMF, 2016, last visited: 2016-05-11. [Online]. Available: https://github.com/atlanmod/NeoEMF
[53] A. Benelallam, A. Gómez, G. Sunyé, M. Tisi, and D. Launay, “Neo4emf, a scalable persistence layer for emf models,” in *Modelling Foundations and Applications*. Springer, 2014, pp. 230–241. [Online]. Available: https://hal.inria.fr/docs/00/96/85/16/PDF/ECMFA2014-Neo4EMF.pdf
[54] “Eclipse memory analyzer (mat) website,” http://www.eclipse.org/mat/, 2015, last visited: 2016-05-1. [Online]. Available: http://www.eclipse.org/mat/
[55] L. Vogel, “Eclipse memory analyzer (mat) - tutorial,” http://www.vogella.com/tutorials/EclipseMemoryAnalyzer/article.html, 2015, last visited: 2016-05-11. [Online]. Available: http://www.vogella.com/tutorials/EclipseMemoryAnalyzer/article.html
[56] K. T. Harald Alfred Weiner, Sven Efftinge, “Eclipse community forums - using neoemf with xtext,” https://www.eclipse.org/forums/index.php/m/1725599/, 2016, last visited: 2016-05-11. [Online]. Available: https://www.eclipse.org/forums/index.php/m/1725599/
[57] H. A. Weiner, “Bug 477617 - missing @since tag,” https://bugs.eclipse.org/bugs/show_bug.cgi?id=477617, 2015, last visited: 2016-05-11 Bug 477617 - missing @since tag. [Online]. Available: https://bugs.eclipse.org/bugs/show_bug.cgi?id=477617
[58] K. T. Harald Alfred Weiner, “Bug 477623 - missing @since tag - genmodel,” https://bugs.eclipse.org/bugs/show_bug.cgi?id=477623, 2015, last visited: 2016-05-11. [Online]. Available: https://bugs.eclipse.org/bugs/show_bug.cgi?id=477623
[59] E. M. Ingo Mohr, “Bug 428088 - @since tags should be generated to both getters and setters,” https://bugs.eclipse.org/bugs/show_bug.cgi?id=428088, 2014, last visited: 2016-05-11. [Online]. Available: https://bugs.eclipse.org/bugs/show_bug.cgi?id=428088
[60] G. D. Harald Alfred Weiner, “Npe and inconsistencies for containment list - issue nr 7,” https://github.com/atlanmod/NeoEMF/issues/7, 2016, last visited: 2016-05-11. [Online]. Available: https://github.com/atlanmod/NeoEMF/issues/7
[61] ——, “commons-io not included in plug-in files - isissue nr 8,” https://github.com/atlanmod/NeoEMF/issues/8, 2016, last visited: 2016-05-11. [Online]. Available: https://github.com/atlanmod/NeoEMF/issues/8