Exploring the Limits of Domain Model Recovery

Paul Klint, Davy Landman, Jurgen Vinju
Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
{Paul.Klint, Davy.Landman, Jurgen.Vinju}@cwi.nl

Abstract—We are interested in re-engineering families of legacy applications towards using Domain-Specific Languages (DSLs). Is it worth to invest in harvesting domain knowledge from the source code of legacy applications?

Reverse engineering domain knowledge from source code is sometimes considered very hard or even impossible. Is it also difficult for “modern legacy systems”? In this paper we select two open-source applications and answer the following research questions: which parts of the domain are implemented by the application, and how much can we manually recover from the source code? To explore these questions, we compare manually recovered domain models to a reference model extracted from domain literature, and measured precision and recall.

The recovered models are accurate: they cover a significant part of the reference model and they do not contain much junk. We conclude that domain knowledge is recoverable from “modern legacy” code and therefore domain model recovery can be a valuable component of a domain re-engineering process.

I. INTRODUCTION

There is ample anecdotal evidence \[1\] that the use of Domain-Specific Languages (DSLs) can significantly improve productivity of software development, especially the maintenance part. DSLs model expected variations in both time (versions) and space (product families) such that some types of maintenance can be done on a higher level of abstraction and with higher levels of reuse. However, the initial investment in designing a DSL can be prohibitively high because a complete understanding of a domain is required. Moreover, when unexpected changes need to be made that were not catered for in the design of the DSL, the maintenance costs can be relatively high. Both issues indicate how both the quality of domain knowledge and the efficiency of acquiring it are pivotal for the success of a DSL based software maintenance strategy.

In this paper we investigate the source code of existing applications as valuable sources of domain knowledge. DSLs are practically never developed in green field situations. We know from experience that rather the opposite is the case: several comparable applications by the same or different authors are often developed before we start considering a DSL. So, when re-engineering a family of systems towards a DSL, there is opportunity to reuse knowledge directly from people, from the documentation, from the user interface (UI) and from the source code. For the current paper we assume the people are no longer available, the documentation is possibly wrong or incomplete and the UI may hide important aspects, so we scope the question to recovering domain knowledge from source code. Is valuable domain knowledge present that can be included in the domain engineering process?

From the field of reverse engineering we know that recovering this kind of design information can be hard \[2\]. Especially for legacy applications written in low level languages, where code is not self-documenting, it may be easier to recover the information by other means. On the other hand, if a legacy application was written in a younger object-oriented language, should we not expect to be able to retrieve valuable information about a domain? This sounds good, but we would like to observe precisely how well domain model recovery from source code could work in reality. Note that both the quality of the recovered information and the position of the observed applications in the domain are important factors.

A. Positioning domain model recovery

One of the main goals of reverse engineering is design recovery \[2\] which aims to recover design abstractions from any available information source. A part of the recovered design is the domain model.

Design recovery is a very broad area, therefore, most research has focused on sub-areas. The concept assignment problem \[3\] tries to both discover human-oriented concepts and connect them to the location in the source code. Often this is further split into concept recovery \[4\]–\[6\], and concept location \[7\]. Concept location, and to a lesser extent concept recovery, has been a very active field of research in the reverse engineering community.

However, the notion of a concept is still very broad and features are an example of narrowed-down concepts and one can identify the sub-areas of feature location \[8\] and feature recovery. Domain model recovery as we will use in this paper is a closely related sub-area. We are interested in a pure domain model, without the additional artifacts introduced by software design and implementation. The location of these artifacts is not interesting either. For the purpose of this paper, a domain model (or model for short) consists of entities and relations between these entities.

Abebe et al.’s \[9, 10\] domain concept extraction is similar to our sub-area. As is Ratiu et al.’s \[11\] domain ontology recovery. In Section \[12\] we will further discuss these relations.

B. Research questions

To learn about the possibilities of domain model recovery we pose this question: how much of a domain model can be recovered under ideal circumstances? By ideal we mean that the applications under investigation should have well-structured and self-documenting object-oriented source code.

\[1\] Also known as concept mining, topic identification, or concept discovery.
This leads to the following research questions:

Q1. Which parts of the domain are implemented by the application?

Q2. Can we manually recover those implemented parts from the object-oriented source code of an application?

Note that we avoid automated recovery here because any inaccuracies introduced by tool support could affect the validity or accuracy of our results.

Figure 1 illustrates the various domains that are involved: The Reference Model (REF) represents all the knowledge about a specific domain and acts as oracle and upper limit for the domain knowledge that can be recovered from any application in that domain. The Recovered Model (REC) is the domain knowledge obtained by inspecting the source code of the application. The Observed Model (OBS) represents the part of the reference domain that an application covers, i.e. all the knowledge about a specific application in the domain that a user may obtain by observing its external behavior and its documentation but not its internal structure.

Ideally, both domain models should completely overlap, however, there could be entities in OBS not present in REC and vice versa. Therefore, figure 2 illustrates the final mapping we have to make, between SRC and USR. The Intra-Application Model (INT) represents the knowledge recovered from the source code, also present in the user view, without limiting it to the knowledge found in REF.

In Section II we describe our research method, explaining how we will analyze the mappings between USR and REF (OBS), SRC and REF (REC), and SRC and USR (INT) in order to answer Q1 and Q2. The results of each step are described in detail in Sections III to VIII. Related work is discussed in Section IX and Section X (Conclusions) completes the paper.

II. RESEARCH METHOD

In order to investigate the limits of domain model recovery we study manually extracted domain models. The following questions guide this investigation:

A) Which domain is suitable for this study?
B) What is the upper limit of domain knowledge, or what is our reference model (REF)?
C) How to select two representative applications?
D) How do we recover domain knowledge that can be observed by the user of the application (Q1 & OBS)?
E) How do we recover domain knowledge from the source code (Q2 & REC)?

F) How do we compare models that use different vocabularies (terms) for the same concepts? (Q1, Q2)?
G) How do we compare the various domain models to measure the success of domain model recovery? (Q1, Q2)?

We will now answer the above questions in turn. Although we are exploring manual domain model recovery, we want to make this manual process as traceable as possible since this enables independent review of our results. Where possible we automate the analysis (calculation of metrics, precision and recall), and further processing (visualization, table generation) of manually extracted information. Both data and automation scripts are available online.

A. Selecting a target domain

We have selected the domain of project planning for this study since it is a well-known, well-described, domain of manageable size for which many open source software applications exist. We use the Project Management Body of Knowledge (PMBOK) published by Project Management Institute (PMI) for standard terminology in the project management domain. Note that as such the PMBOK covers a lot more than just project planning.

B. Obtaining the Reference Model (REF)

Validating the results of a reverse engineering process is difficult and requires an oracle, i.e., an actionable domain model suitable for comparison and measurement. We have transformed the descriptive knowledge in PMBOK into such a reference model using the following, traceable, process:

1) Read the PMBOK book.
2) Extract project planning facts.
3) Assign a number to each fact and store its source page.
4) Construct a domain model, where each entity, attribute, and relation are linked to one or more of the facts.
5) Assess the resulting model and repeat the previous steps when necessary.

The resulting domain model will act as our Reference Model.

and Section III gives the details.

C. Application selection

In order to avoid bias towards a single application, we need at least two project planning applications to extract domain models from. Section IV describes the selection criteria and the selected applications.

2See http://homepages.cwi.nl/~landman/icsm2013/
D. Observing the application

A user can observe an application in several ways, ranging from its UI, command-line interface, configuration files, documentation, scripting facilities and other functionality or information exposed to the user of the application. In this study we use the UI and documentation as proxies for what the user can observe.

We have followed these steps to obtain the User Model (USR) of the application:

1) Read the documentation.
2) Determine use cases.
3) Run the application.
4) Traverse the UI depth-first for all the use cases.
5) Collect information about the model exposed in the UI.
6) Construct a domain model, where each entity and relation are linked to a UI element of the application.
7) Assess the resulting model and repeat the previous steps when necessary.

We report about the outcome in Section V.

E. Inspecting the source code

We have designed the following traceable process to extract a domain model from each application’s source code, the Source Model (SRC):

1) Read the source code as if it is plain text.
2) Extract project planning facts.
3) Store its filename, and line number (source location).
4) Construct a model, where each entity, attribute, and relation is linked to a source location in the application’s source code.
5) Assess the model and repeat the previous steps when necessary.

The results appear in Section VI.

F. Mapping models

After performing the above steps we have obtained five domain models for the same domain, derived from different sources:

- The Reference Model (REF) derived from PMBOK.
- For each of the two applications:
  - User Model (USR).
  - Source Model (SRC).

While all these model are in the project planning domain, they all use different vocabularies. Therefore, we have to manually map the models to the same vocabulary. Mapping the USR and SRC models onto the REF model, gives the Observed (OBS) and Recovered Model (REC).

The final mapping we have to make, is between the SRC and USR models. We want to understand how much of the User Model (USR) is present in the Source Model (SRC). Therefore, we also map the SRC onto the USR model, giving the Intra-Application Model (INT). The results of all these mappings are given in Section VII.

G. Comparing models

To be able to answer Q1 and Q2, we will compare the 11 produced models. Following other research in the field of concept assignment, we use the most common information retrieval (IR) approach, recall and precision, for measuring quality of the recovered data. Recall measures how much of the expected model is present in the found model, and precision measures how much of the found model is part of the expected.

We have designed the following traceable process to extract a domain model from each application’s source code, the Source Model (SRC):

1) Read the source code as if it is plain text.
2) Extract project planning facts.
3) Store its filename, and line number (source location).
4) Construct a model, where each entity, attribute, and relation is linked to a source location in the application’s source code.
5) Assess the model and repeat the previous steps when necessary.

The results appear in Section VI.

The final mapping we have to make, is between the SRC and USR models. We want to understand how much of the User Model (USR) is present in the Source Model (SRC). Therefore, we also map the SRC onto the USR model, giving the Intra-Application Model (INT). The results of all these mappings are given in Section VII.

III. PROJECT PLANNING REFERENCE MODEL

Since there is no known domain model or ontology for project planning that we are aware of, we need to construct one ourselves. The aforementioned PMBOK [12] is our point of departure. PMBOK avoids project management style specific terminology, making it well-suited for our information needs.

A. Gathering facts

We have analyzed the whole PMBOK book. This analysis has been focused on the concept of a project and everything related to project planning therefore we exclude other concepts and processes in the project management domain.

After analyzing 467 pages we have extracted 151 distinct facts related to project planning. A fact is either an explicitly defined concept, an implicitly defined concept based on a summarized paragraph, or a relations between concepts. These facts were located on 67 different pages. This illustrates that project planning is a subdomain and that project management as a whole covers many topics that fall outside the scope of the current paper. Each fact was assigned a unique number and the source page number where it was found in the current paper. Each fact was assigned a unique number and the source page number where it was found in the current paper. Each fact was assigned a unique number and the source page number where it was found in the current paper.

Two example facts are: “A milestone is a significant point or event in the project.” (id:108, page: 136) and “A milestone may be mandatory or optional.” (id:109, page: 136).

B. Creating the Reference Model REF

In order to turn these extracted facts into a model for project planning, we have translated the facts to entities, attributes

Implementation details or concepts from other domains.
TABLE I
NUMBER OF ENTITIES AND RELATIONS IN THE CREATED MODELS, AND THE AMOUNT OF LOCATIONS IN THE PMBOK BOOK, SOURCE CODE, OR UI SCREENS USED TO CONSTRUCT THE MODEL.

<table>
<thead>
<tr>
<th>Source</th>
<th># entities</th>
<th># relations</th>
<th>unique observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>PMBOK REF</td>
<td>74</td>
<td>75</td>
<td>32</td>
</tr>
<tr>
<td>Endeavour USR</td>
<td>23</td>
<td>30</td>
<td>8</td>
</tr>
<tr>
<td>SRC</td>
<td>26</td>
<td>51</td>
<td>8</td>
</tr>
<tr>
<td>OpenPM USR</td>
<td>22</td>
<td>24</td>
<td>3</td>
</tr>
<tr>
<td>SRC</td>
<td>28</td>
<td>44</td>
<td>6</td>
</tr>
</tbody>
</table>

Fig. 3. Fragment of reference model REF visualized as class diagram. Boxes represent entities, arrows relations, and dashed lines link entities to relations.

of entities, and relations between entities. The two example facts (108 and 109), are translated into a relation between the classes Project and Milestone, and the mandatory attribute for the Milestone class. The meta-model of our domain model is a class diagram. We use a textual representation in the meta-programming language Rascal [13] which is also used to perform calculations on these models (precision, recall).

Table II characterizes the size of the project planning reference domain model REF by number of entities, relations and attributes; it contains of 74 entities and 107 relations. There is also a set of 49 attributes, but this seems incomplete, because in general we expect any entity to have more than one property. The lack of details in PMBOK could be an explanation for this. Therefore, we did not use the attributes of the reference model to calculate similarity.

The model is too large to include in this paper, however for demonstration purposes, a small subset is shown in Figure 3.

Not all the facts extracted from PMBOK are used in the Reference Model. Some facts carry only explanations. For example “costs are the monetary resources needed to complete the project”. Some facts explain dynamic relations that are not relevant for an entity/relationship model. These two categories explain 55 of the 68 unused facts. The remaining 13 facts were not clear enough to be used or categorized. In total 83 of the 151 observed facts are represented in the Reference Model.

C. Discussion

We have created a Reference Model that can be used as oracle for domain model recovery and other related reverse engineering tasks in the project planning domain. The model was created by hand by the second author, and care was taken to make the whole process traceable. We believe this model can be used for other purposes in this domain as well, such as application comparison and checking feature completeness.

Threats to validity: We use PMBOK as main source of information for project planning. There are different approaches to project planning and a potential threat is that some are not covered in this book. Since PMBOK is an industry standard (ANSI and IEEE), we consider this to be a low-risk threat and have not mitigated it.

Another threat is that model recovery by another person could lead to a different model. The traceable extraction of the reference model makes it possible to understand the decisions on which the differences are based. Due to the availability of our analysis scripts, the impact of differences can be easily computed.

IV. APPLICATION SELECTION

We are interested in finding “ideal” project planning systems to manually read and extract domain models from. The following requirements have guided our search:

- Source code is available: to enable analysis at all.
- No more than 30 KSLOC: to keep manual analysis feasible.
- Uses an explicit data model, for example Model View Controller (MVC), or an Object-relational mapping (ORM): to ensure that domain elements can be identified in the source code.

We have made a shortlist of 10 open source project planning systems [14]. The list contains applications implemented in different languages (Java, Ruby, and C++) and sizes ranging from 18 KSLOC to 473 KSLOC.

From this Endeavour and OpenPM satisfy the aforementioned requirements. Endeavour is a Java application that uses a custom MVC design with ThinWire as front-end framework, and Hibernate as ORM. OpenPM uses Java servlets in combination with custom JavaScript. It also uses Hibernate as ORM. Table II and III describe the structure and size of the two applications. Note that OpenPM’s view package contained MVC controller logic, and the servlets the MVC views.

Both systems aim at supporting the process of planning by storing the process state but they hardly support process enforcement, except recording dependence between activities.

A. Discussion

A threat to external validity is that both systems are implemented in Java. Looking at systems in multiple modern languages is considered future work.

V. OBTAINING THE USER MODEL

We have used the UI and documentation of the applications to construct the User Model (USR). Use cases were extracted from the documentation when possible [15]. Following these use cases, a depth-first exploration of the UI is performed. For

4 ChilliProject, Endeavour, GanttProject, LibrePlan, OpenPM, OpenProj, PLANdora, project.net, taskjuggler, Xplanner+

5 Number of files and SLOC are calculated using the cloc tool [14].

6 Unfortunately, OpenPM does not provide documentation.
We have chosen the Eclipse Integrated Development Environment (IDE) to read the source code of the selected applications. Our goal was to maximize the amount of information we could recover. Therefore, we have first read the source code and then used Rascal to analyze relations in the source code. Rascal uses Eclipse’s JDT to analyze Java code, and provides a visualization library that can be used to quickly verify hypothesis formed during the first read-through.

For the actual creation of the model, we have designed and followed these rules:

- Read only the source code, not the database scheme/data.
- Do not run the application.
- Use the terms of the applications, do not translate them to terms used in the Reference Model.
- Include the whole model as seen by the application, do not filter out obvious implementation entities.
- Do read comments and string literals.

We have used the same meta-model as used for describing the Reference Model. We replaced the fact’s identifiers with source locations (filename and character range), which are a native construct in Rascal. To support the process of collecting facts from the source code we added a menu-item to the context-menu of the Java editor to write the cursor’s source location to the clipboard.

The domain model for each application was created in a similar fashion as we did when creating the reference model. All the elements in the domain model are based on one or more specific observations in the source code (see table I).

For example the relation between Task and Dependency in Endeavour’s SRC model is based on the List<Dependency> dependencies field found on line 35 in file Endeavour-Mgmt/model/org/endeavour/mgmt/model/Task.java.

### B. Results

Table II shows the sizes of the extracted models for both applications expressed in number of entities, relations and attributes and the number of unique source code locations where they were found.

1) **Endeavour**: In Endeavour 26 files contributed to the domain model. 22 of those files were in the model package, the other 4 were from the controller package. The controller classes each contributed one fact. 155 of the source locations were from the model package.

2) **OpenPM**: In OpenPM 22 files contributed to the domain model. These files were all located in the model package.

### C. Discussion

We have performed domain model recovery on two open source software applications for project planning.

Both applications use the same ORM system, but a different version of the API. Endeavour also contains a separate view model, which is used in the MVC user interface. However, it has been implemented as a pass-through layer for the real model.

**Threats to validity**: A first threat (to internal validity) is that manual analysis is always subject to bias from the performer and that this was performed by the same author who created the other models. We have mitigated this by maximizing the traceability of our analysis: we have followed a fixed analysis...
process and have performed multiple analysis passes over the source code and published the data.

A second threat (to external validity) is the limited size of the analyzed applications, both contain less than 20 KSLOC Java. Larger applications would make our conclusions more interesting and general, but they would also make the manual analysis less feasible.

VII. MAPPING MODELS

We now have five domain models of project planning: one reference model (REF) to be used as oracle, and four domain models (SRC, USR) obtained from the two selected project planning applications. These models use different vocabulary, we have to map them onto the same vocabulary to be able to compare them.

A. Lightweight domain model mapping

We manually map the entities between different comparable models. The question is how to decide whether to entities are the same. Strict string equality is too limited and should be relaxed to some extent.

Table [IV] and [V] show the mapping categories we have identified for the (un)successful mapping of model entities.

B. Mapping results

We have manually mapped all the entities in the User Model (USR) and the Source Model (SRC) to the Reference Model (REF), and SRC to USR. For each mapping we have explicitly documented the reason for choosing this mapping. For example, in Endeavour’s SRC model the entity Iteration is mapped to Milestone in the Reference Model using specialization, with documented reason: “Iterations split the project into chunks of work, Milestones do the same but are not necessarily iterative.”

Table [VI] and [VII] contain the number of mapping categories used for both applications, per mapping. For some mapping categories, it is possible for one entity to map to multiple, or multiple entities to one. For example the Task and WorkProduct entities in Endeavour’s SRC model are mapped on the Activity entity in the Reference Model. Therefore, we report the numbers of the entities in both the models, the source and the target.

The relatively large number of identically named entities (7/15) between Endeavour and the reference model is due to the presence of a similar structure of five entities, describing all the possible activity dependencies.

An example of a failed mapping is the ObjectVersion entity in the Source Model of OpenPM. This entity is an implementation detail. It is a variant of the Temporal Object pattern\footnote{See \url{http://martinfowler.com/eaaDev/TemporalObject.html}} where

\begin{table}[h]
\centering
\caption{Categories for successfully mapped entities}
\begin{tabular}{ll}
\hline
Mapping name & Description \\
\hline
Equal Name & Entity has the same name as an entity in the other model. Note that this is the only category which can also be a failure when the same name is used for semantically different entities. \\
Synonym & Entity is a direct synonym for an entity in the other model, and is it not a homonym. \\
Extension & Entity captures a wider concept than the same entity in the other model. \\
Specialization & Entity is a specific or concrete instance of the same entity in the other model. \\
Implementation specialization & Comparable to specialization but the specialization is related to an implementation choice. \\
\hline
\end{tabular}
\end{table}

\begin{table}[h]
\centering
\caption{Categories for unsuccessfully mapped entities}
\begin{tabular}{ll}
\hline
Mapping name & Description \\
\hline
Missing & The domain entity is missing in the other model, i.e. a false positive. This is the default mapping failure when an entity cannot be mapped via any of the other categories. \\
Implementation & The entity is an implementation detail and is not a real domain model entity. \\
Too detailed & An entity is a domain entity but is too detailed in comparison with the other model. \\
Domain detail & The entity is a detail of a sub domain, this category is a subclass of "too detailed". \\
\hline
\end{tabular}
\end{table}

\begin{table}[h]
\centering
\caption{Endeavour: Entities in the mapped models, per mapping category}
\begin{tabular}{lllllllllll}
\hline
Category & USR & REF & SRC & REF & SRC & USR \\
\hline
Equal Name & 7 & 7 & 7 & 7 & 21 & 21 \\
Synonym & 2 & 3 & 2 & 3 & 3 & 2 \\
Extension & 0 & 0 & 0 & 0 & 0 & 0 \\
Specialization & 5 & 3 & 5 & 3 & 0 & 0 \\
Implementation specialization & 1 & 1 & 1 & 1 & 0 & 0 \\
Total & 15 & 14 & 15 & 14 & 24 & 23 \\
\hline
Equal Name & 1 & - & 1 & - & 0 & - \\
Missing & 1 & - & 2 & - & 0 & - \\
Implementation & 1 & - & 2 & - & 2 & 0 \\
Domain Detail & 5 & - & 6 & - & 0 & - \\
Too Detailed & 0 & - & 0 & - & 0 & - \\
Total & 8 & - & 11 & - & 2 & - \\
\hline
\end{tabular}
\end{table}

\begin{table}[h]
\centering
\caption{OpenPM: Entities in mapped models, per mapping category}
\begin{tabular}{lllllllllll}
\hline
Category & USR & REF & SRC & REF & SRC & USR \\
\hline
Equal Name & 1 & 1 & 1 & 1 & 18 & 18 \\
Synonym & 3 & 3 & 4 & 4 & 4 & 4 \\
Extension & 1 & 1 & 1 & 1 & 0 & 0 \\
Specialization & 0 & 0 & 0 & 0 & 0 & 0 \\
Implementation specialization & 1 & 1 & 1 & 1 & 0 & 0 \\
Total & 6 & 6 & 7 & 7 & 22 & 22 \\
\hline
Missing & 2 & - & 2 & - & 1 & - \\
Implementation & 12 & - & 17 & - & 5 & - \\
Domain Detail & 0 & - & 0 & - & 0 & - \\
Too Detailed & 2 & - & 2 & - & 0 & - \\
Total & 16 & - & 21 & - & 6 & - \\
\hline
\end{tabular}
\end{table}
We have used a lightweight approach for mapping domain modified all previous classifications have been reconsidered. During the process and each time a category was added or two applications, however we have no guarantees for other categories presented above, turned out to be sufficient for these categories. The categories have evolved during the process and each time a category was added or modified all previous classifications have been reconsidered.

We now have five manually constructed and six derived domain models for project planning:

- One reference model (REF) to be used as oracle.
- Four domain models (SRC, USR) obtained from each of the two selected project planning applications.
- Six derived domain models (OBS, REC, INT) resulting from the mapping of the previous four (SRC, USR).

How can we compare these models in a meaningful way?

A. Recall and Precision

The most common measures to compare the results of an IR technique are recall and precision. Often it is not possible to get the 100% in both, and we have to discuss which measure is more important in the case of our model comparisons.

We have more than two datasets, and depending on the combination of datasets, recall or precision is more important. Table IX explains in detail how recall and precision will be used and explains for the relevant model combinations which measure is useful and what will be measured.

Given two models $M_1$ and $M_2$, we use the following notation. The comparison of two models is denoted by $M_1 \circ M_2$ and results in recall and precision for the two models. If needed, $M_1$ is first mapped to $M_2$ as described in Tables VII and VII.

B. Results

Tables X and XI shows the results for, respectively, Endeavour and OpenPM. Which measures are calculated is based on the analysis in Table IX.

C. Relation Similarity

Since recall and precision for sets of entities provides no insight into similarity of the relations between entities, we need an additional measure. Our domain models contain entities and their relations. Entities represent the concepts of the domain, and relations their structure. If we consider the relations as a set of edges, we can directly calculate recall and precision in a similar fashion as described above.

We also considered some more fine grained metrics for structural similarity. Our domain model is equivalent to a subset of Unified Modeling Language (UML) class diagrams and several approaches exist for calculating the minimal difference between such diagrams [15, 16]. Such “edit distance” methods give precise indications of how big the difference is. Similarly we might use general graph distance metrics [17]. We tried this latter method and found that the results, however more sophisticated, were harder to interpret. For example, USR and REF were 11% similar for Endeavor. This seems to be in line with the recall numbers, 6% for relations and 19% for entities, but the interesting precision results (64% and 15%) are lost in this metric. So we decided not to report these results and stay with the standard accuracy analysis.

<table>
<thead>
<tr>
<th>Source</th>
<th>Model Entities†</th>
</tr>
</thead>
<tbody>
<tr>
<td>PMBOK</td>
<td>Reference Model</td>
</tr>
<tr>
<td>END</td>
<td></td>
</tr>
<tr>
<td>USR</td>
<td></td>
</tr>
<tr>
<td>SRC</td>
<td></td>
</tr>
<tr>
<td>OpenPM</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>Model Entities†</th>
</tr>
</thead>
<tbody>
<tr>
<td>PMBOK</td>
<td>Reference Model</td>
</tr>
<tr>
<td>END</td>
<td></td>
</tr>
<tr>
<td>USR</td>
<td></td>
</tr>
<tr>
<td>SRC</td>
<td></td>
</tr>
<tr>
<td>OpenPM</td>
<td></td>
</tr>
</tbody>
</table>

Bold entity in Reference Model is used in application models. Bold entity in application model could be mapped to entity in Reference Model.

### TABLE VIII

<table>
<thead>
<tr>
<th>Entities found in the various domain models.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
</tr>
<tr>
<td>--------</td>
</tr>
<tr>
<td>PMBOK</td>
</tr>
<tr>
<td>END</td>
</tr>
<tr>
<td>USR</td>
</tr>
<tr>
<td>SRC</td>
</tr>
<tr>
<td>OpenPM</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>Model Entities†</th>
</tr>
</thead>
<tbody>
<tr>
<td>PMBOK</td>
<td>Reference Model</td>
</tr>
<tr>
<td>END</td>
<td></td>
</tr>
<tr>
<td>USR</td>
<td></td>
</tr>
<tr>
<td>SRC</td>
<td></td>
</tr>
<tr>
<td>OpenPM</td>
<td></td>
</tr>
</tbody>
</table>

Bold entity in Reference Model is used in application models. Bold entity in application model could be mapped to entity in Reference Model.

### VIII. Comparing the models

We now have five manually constructed and six derived domain models for project planning:

- One reference model (REF) to be used as oracle.
- Four domain models (SRC, USR) obtained from each of the two selected project planning applications.
- Six derived domain models (OBS, REC, INT) resulting from the mapping of the previous four (SRC, USR).

How can we compare these models in a meaningful way?

A. Recall and Precision

The most common measures to compare the results of an IR technique are recall and precision. Often it is not possible to get the 100% in both, and we have to discuss which measure is more important in the case of our model comparisons.

We have more than two datasets, and depending on the combination of datasets, recall or precision is more important. Table IX explains in detail how recall and precision will be used and explains for the relevant model combinations which measure is useful and what will be measured.

Given two models $M_1$ and $M_2$, we use the following notation. The comparison of two models is denoted by $M_1 \circ M_2$ and results in recall and precision for the two models. If needed, $M_1$ is first mapped to $M_2$ as described in Tables VII and VII.

B. Results

Tables X and XI shows the results for, respectively, Endeavour and OpenPM. Which measures are calculated is based on the analysis in Table IX.

C. Relation Similarity

Since recall and precision for sets of entities provides no insight into similarity of the relations between entities, we need an additional measure. Our domain models contain entities and their relations. Entities represent the concepts of the domain, and relations their structure. If we consider the relations as a set of edges, we can directly calculate recall and precision in a similar fashion as described above.

We also considered some more fine grained metrics for structural similarity. Our domain model is equivalent to a subset of Unified Modeling Language (UML) class diagrams and several approaches exist for calculating the minimal difference between such diagrams [15, 16]. Such “edit distance” methods give precise indications of how big the difference is. Similarly we might use general graph distance metrics [17]. We tried this latter method and found that the results, however more sophisticated, were harder to interpret. For example, USR and REF were 11% similar for Endeavor. This seems to be in line with the recall numbers, 6% for relations and 19% for entities, but the interesting precision results (64% and 15%) are lost in this metric. So we decided not to report these results and stay with the standard accuracy analysis.
D. Discussion

1) Low precision and recall for relations: On the whole the results for the precision and recall of the relation part of the models are lower than the quality of the entity mappings. We investigated this by taking a number of samples. The reason is that the Reference model is more detailed, introducing intermediate entities with associated relations. For every intermediate entity, two or more relations are introduced which can not be found in the recovered models.

These results indicate that the recall and precision metrics for sets of relations underestimate the structural similarity of the models.

2) Precision of OBS: USR $\triangleright$ REF: We found the precision of OBS to be 64% (Endeavour) and 23% (OpenPM), indicating that both applications contain a significant amount of entities that are unrelated to project planning as delimited by the Reference Model. For Endeavour, out of the 8 unmappable entities (see Table VII in section VII), only 2 were actual implementation details. The other 6 are sub-domain details not globally shared within the domain. If we recalculate to correct for this, Endeavour’s Observed Model even has a precision of 91%. For OpenPM there are only 2 out of the 16 for which this correction can be applied, leaving the precision at 36%. For the best scenario, in this case represented by Endeavour, 90% of the User Model (USR) is part of the Reference Model (REF).

OpenPM’s relatively low precision (36%) can be explained by Table VII which show the USR model has a lot of implementation detail (related to version control operations).

3) Recall of OBS: USR $\triangleright$ REF: The recall for the Observed Model (OBS) is for Endeavour 19% and for OpenPM 7%. Which means both applications cover less than 20% of the project planning domain.

4) Precision of REC: SRC $\triangleright$ REF: The precision of the Recovered Model (REC) is for Endeavour 56% (corrected 88%), and for OpenPM 25% (corrected 39%). This shows that for the best scenario, represented again by Endeavour, the Source Model only contains 12% implementation details.

5) Recall of REC: SRC $\triangleright$ REF: The recall for the Recovered Model (REC) is for Endeavour 19% and for OpenPM 9%. The higher recall for OpenPM, compared to OBS, for both entities and relations is an example where the Source Model contained more information then the User Model, which we will discuss in the next paragraph.

6) Precision and recall for INT: SRC $\triangleright$ USR: How much of the User Model can be recovered by analyzing only the Source Model? For both Endeavour and OpenPM, recall is 100%. This means that every entity in the USR model was found in the source code. Endeavour’s precision was 92% and OpenPM’s 79%. OpenPM contains an example where information in the Source Model is not observable in the User Model: comments in the source code explain the Milestones and their relation to Iterations.

The 100% recall and high precision mean that these applications were indeed amenable for reverse engineering (as we hypothesized when selecting these applications). We could extract most of the information from the source code. For this comparison, even the relations score quite high. This indicates that User Model and Source Model are structurally similar. Manual inspection of the models confirm this.

7) Recall for Endeavour and OpenPM combined: Endeavour’s and OpenPM’s recall of USR $\triangleright$ REF and SRC $\triangleright$ REF measure the coverage of the domain a re-engineer can achieve. How much will the recall improve if we combine the recovered models of the two systems?
We only have two small systems, however, Table XII contains the recall and precision for Endeavour and OpenPM combined. A small increase in recall, from 19% to 23%, indicates that there is a possibility for increasing the recall by observing more systems. However, as expected, at the cost of precision.

8) Interpretation: Since our models are relatively small, our results cannot be statistically significant but are only indicative. Therefore we should not report exact percentages, but characterizing our recall and precision as high seems valid. Further research based on more applications is needed to confirm our results.

IX. RELATED WORK

There are many connections between ontologies and domain models. The model mappings that we need are more specific than the ones provided by general ontology mapping [18].

Abebe and Tonella [9] introduced a natural language parsing (NLP) method for extracting an ontology from source code. They came to the same conclusion as we do: this extracted ontology contains a lot of implementation details. Therefore, they introduced an IR filtering method [10] but it was not as effective as the authors expected. Manual filtering of the IR keyword database was shown to improve effectiveness. Their work is in the same line as ours, but we have a larger reference domain model, and we focus on finding the limits of domain model recovery, not on an automatic approach. It would be interesting to apply their IR filtering to our extracted models.

Rauti et al. [11] proposed an approach for domain ontology extraction. Using a set of translation rules they extract domain knowledge from the API of a set of related software libraries. Again, our focus is on finding the limits of model recovery, not on automating the extraction.

Hsi et al. [19] introduced ontology excavation. Their methodology consists of a manual depth-first modeling of all UI interactions, and then manually creating an ontology, filtering out non-domain concepts. They use five graph metrics to identify interesting concepts and clusters in this domain ontology. We are interested in finding the domain model inside the user-interface model, Hsi et al. perform this filtering manually, and then look at the remaining model. Automatic feature extraction of user interfaces is described in [20].

Carey and Gannod [6] introduced a method for concept identification. Classes are considered the lowest level of information of an object-oriented system and Machine Learning is used in combination with a set of class metrics. This determines interesting classes, which should relate to domain concepts. Our work is similar, but we focus on all the information in the source code, and are interested in the maximum that can be recovered from the source. It could be interesting to use our reference model to measure how accurately their approach removes implementation concerns.

UML class diagram recovery [21], [22] is also related to our work but has a different focus. Research focuses on the precision of the recovered class diagrams, for example the difference between a composition and aggregation relation. We are interested in less precise UML class diagrams.

Work on recovering the concepts, or topics, of a software system [4], [5] has a similar goal as ours. IR techniques are used to analyze all the terms in the source code of a software system, and find relations or clusters. Kuhn et al. [5] use identifiers in source code to extract semantic meaning and report on the difficulty of evaluating their results. Our work focuses less on structure and grouping of concepts and we evaluate our results using a constructed reference model.

Reverse engineering the relation between concepts or features [8], [23], assumes that there is a set of known features or concepts and tries to recover the relations between them. These approaches are related to our work since the second half of our problem is similar: after we have recovered domain entities, we need to understand their relations.

DeBaud et al. [24] report on a domain model recovery case study on a COBOL program. By manual inspection of the source code, a developer reconstructed the data constructs of the program. They also report that implementation details make extraction difficult, and remark that systems often implement multiple domains, and that the implementation language plays an important role in the discovery of meaning in source code.

We do not further discuss other related work on knowledge recovery that aims at extracting facts about architecture or implementation. One general observation in all the cited work is that it is hard to separate domain knowledge from implementation knowledge.

X. CONCLUSIONS

We have explored the limits of domain model recovery via a case study in the project planning domain. Here are our results and conclusions.

A. Reference model

Starting with PMBOK as authoritative domain reference we have manually constructed an actionable domain model for project planning. This model is openly available and may be used for other reverse engineering research projects.

B. Lightweight model mapping

Before we can understand the differences between models, we have to make them comparable by mapping them to a common model. We have created a manual mapping method that determines for each entity if and how it maps onto the target model. The mapping categories evolved while creating the mappings. We have used this approach to describe six useful mappings, four to the Reference Model and two to the User Model.

C. What are the limits of domain model recovery?

We have formulated two research questions to get insight in the limits of domain model recovery. Here are the answers we have found (also see Table IX and remember our earlier comments on the interpretation of the percentages given below).

Q1: Which parts of the domain are implemented by the application? Using the user view (USR) as a representation of the part of the domain that is implemented by an application, we have created two domain models for each of the two selected
applications. These domain models represent the domain as exposed by the application. Using our Reference Model (REF) we were able to determine which part of USR was related to project planning. For our two cases 91% and 36% of the User Model (USR) can be mapped to the Reference Model (REF). This means 9% and 64% of the UI is about topics not related to the domain. From the user perspective we could determine that the applications implement 19% and 7% of the domain.

The tight relation between the USR and the SRC model (100% recall) shows us that this information is indeed explicit and recoverable from the source code. Interestingly, some domain concepts were found in the source code that were hidden by the UI and the documentation, since for OpenPM the recall between USR and REF was 7% where it was 9% between SRC and REF.

So, the answer for Q1 is: the recovered models from source code are useful, and only a small part of the domain is implemented by these tools (only 7-19%).

Q2: Can we recover those implemented parts from the source of the application? Yes, see the answer to Q1. The high recall between USR and SRC shows that the source code of these two applications explicitly models parts of the domain. The high precisions (92% and 79%) also show that it was feasible to filter implementation junk manually from these applications from the domain model.

D. Perspective

For this research we manually recovered domain models from source code to understand how much valuable domain knowledge is present in source code. We have identified several follow-up questions:

- How does the quality of extracted models grow with the size and number of applications studied? (Table XII)
- How can differences and commonalities between applications in the same domain be mined to understand the domain better?
- How does the quality of extracted models differ between different domains, different architecture/designs, different domain engineers?
- How can the extraction of a User Model help domain model recovery in general. Although we have not formally measured the effort for model extraction, we have noticed that extracting a User Model requires much less effort than extracting a Source Model.
- How do our manually extracted models compare with automatically inferred models?
- What tool support is possible for (semi-)automatic model extraction?
- How can domain models guide the design of a DSL?

Our results of manually extracting domain models are encouraging. They suggest that when re-engineering a family of object-oriented applications to a DSL their source code is a valuable and trustworthy source of domain knowledge, even if they only implement a small part of the domain.

REFERENCES