Software (re)modularization: Fight against the structure erosion and migration preparation
Nicolas Anquetil, Simon Denier, Stéphane Ducasse, Jannik Laval, Damien Pollet, Roland Ducournau, Rodolphe Giroudeau, Marianne Huchard, Jean-Claude König, Abdelhak-Djamel Seriai

To cite this version:
Nicolas Anquetil, Simon Denier, Stéphane Ducasse, Jannik Laval, Damien Pollet, et al.. Software (re)modularization: Fight against the structure erosion and migration preparation. 2010. <lirmm-00534901>

HAL Id: lirmm-00534901
https://hal-lirmm.cnrs.fr/lirmm-00534901
Submitted on 10 Nov 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Software (re)modularization: 
Fight against the structure erosion and migration preparation

N. Anquetil, S. Denier, S. Ducasse, J. Laval, D. Pollet
RMoD INRIA
stephane.ducasse@inria.fr

R. Ducournau, R. Giroudeau, M. Huchard, J.C. König, D. Seriai
LIRMM - CNRS UMR 5506 - Université de Montpellier II - Montpellier (France)
huchard@lirmm.fr

I. CONTEXT

Software systems, and in particular, Object-Oriented systems are models of the real world that manipulate representations of its entities through models of its processes. The real world is not static: new laws are created, concurrents offer new functionalities, users have renewed expectation toward what a computer should offer them, memory constraints are added, etc. As a result, software systems must be continuously updated or face the risk of becoming gradually out-dated and irrelevant [34]. In the meantime, details and multiple abstraction levels result in a high level of complexity, and completely analyzing real software systems is impractical. For example, the Windows operating system consists of more than 60 millions lines of code (500,000 pages printed double-face, about 16 times the Encyclopedia Universalis). Maintaining such large applications is a trade-off between having to change a model that nobody can understand in details and limiting the impact of possible changes. Beyond maintenance, a good structure gives to the software systems good qualities for migration towards modern paradigms as web services or components, and the problem of architecture extraction is very close to the classical remodularization problem.

II. A SIGNIFICANT PROBLEM

Most of the effort while developing and maintaining a software system is spent in supporting its evolution [50]. It is well-known that up to 80% of the total cost of software development project is spent in maintenance and evolution of existing applications [15], [19]. Large IT applications including applications running central and critical business (e.g., command tracking, banking, railway) have to run and evolve over decades.

Such situations are crystallized in the following two laws of Lehman and Belady. These laws, known as the laws of software evolution, stress the fact that software must continuously evolve to stay useful and that this evolution is accompanied by an increase of complexity.

Continuous Changes. “an E-type program—i.e., a software system that solves a problem or implements a computer application in the real world—that is used must be continually adapted else it becomes progressively less satisfactory” [34]

Increasing Complexity. “As a program is evolved its complexity increases unless work is done to maintain or reduce it.”[34]

From a business perspective, maintenance is mandatory and vital. It is worth to note that the Year 2000 Bug was also revealing some business occasions. For example in France SOPRA which was mainly an SSII, has since the year 2000 developed its TMA (Tierce Maintenance Applicative) branch to get into the maintenance business. SOPRA TMA now represents 2500 developers working on maintenance on a total of 11000 developers.

If many organizations use third party software (e.g. COTS) which they don’t have to maintain, this is less true for their core business applications where their market differential must be implemented in the applications that help running day to day activities. As highlighted by the law of Lehman and Belady, these in-house applications must evolve and in this process will extend far beyond their initial structure, in many independent (and sometimes even conflicting) directions. After some time, it becomes mandatory to restructure the application to federate these evolutions inside a more general and renewed architecture to foster possible future evolutions.

Supporting evolution and prepare migration of applications will be always mandatory. Different programming paradigms have been invented to cope with changes: late-binding, the cornerstone of object-oriented programming, is a typical illustration. But we think that no paradigm will eradicate the need for evolution and changes and that the only possible approach is to guide evolution or to repair the damages caused by an inevitable erosion.

III. THE FAILURE OF CURRENT SOLUTIONS

Whereas software (re-)modularization is a relatively old research field in the context of C or Cobol, it is still really important and requires innovative approaches to deal with the complexity of modern systems especially those
developed in OOP languages. Our analysis — also confirmed in the recent literature (e.g., [1], [7], [47]) — is that this failure is a direct consequence of: (1) the complexity of the manipulated concepts and the variety of modular abstractions (subsystems, packages, classes, class hierarchies, late-binding, aspects, various import relationships,...) as well as (2) a monolithic approach: the use of one type of algorithms (clustering) and one kind of system representation (software components interactions found in the source code).

A. Modular Abstractions

Modular constructs have been the focus of a large body of research. Here we give a non-exhaustive list. A lot of work is currently underway in the context of aspect-oriented programming. Module and package systems have been the focus on a large amount of work. The recent work on Units [22], Jiazz [42], Mixjuice [28], MJ [14], JAM [52], mixin layers [48] shows that this topic is a crucial research area. Bergel et al. conducted a survey on modular language constructs that reflects such a diversity [6].

Modular constructs have also been considered at a lower level than classes, e.g., with mixins [9], traits [17]. These constructs denote, here, sets of properties that somewhat represent the differentia in the Aristotelian genus-differentia definition. Overall, remodularization must address both modular construct levels, by clustering related properties for defining classes and related classes for defining packages. Recursively nested class models have also been proposed [20], [44]—however they cannot be considered as long as the point with non-recursive models is not fixed.

Component-based software approach proposed to build software systems by assembling prefabricated reusable components [54]. Assembly consists in connecting matching interfaces of the components: in a connection, a required interface (describing the services a component needs) is connected to a provided interface (describing the services another component offers). Extracting a component architecture from an object-oriented software has many in common with remodularization because classes need to be grouped based on their dependencies to form the components [11].

Current Issues — The class notion is the only universally accepted modular abstraction. Higher and lower level abstractions are often still in flux. Different languages use different concepts such as modules, packages, namespaces. Hence, current issues involve both identifying adequate abstractions and adapting remodularization algorithms to the various alternative abstractions. Moreover, assessing the modularity of software requires specific tools (e.g., metrics, visualization) that must be adapted to each modular abstraction.

B. Remodularization Approaches

Class hierarchy analysis: Class hierarchy analysis has been largely investigated, for restructuring purposes or finding separate concerns (aspects, traits). As it has been shown in [26], most of the refactoring approaches use explicitly or implicitly substructures of those obtained by Formal Concept Analysis (FCA).

The application of clustering algorithms for software remodularization has been intensively studied [2]. Thousands of experiments were conducted to compare different clustering algorithms, different representation schemes and different coupling metrics between files. Although the experiments used procedural systems, many conclusions may be applied to OO systems as well. The extraction of class views based on Formal Concept Analysis has been proposed in [3]. They evaluated how FCA supports the identification of traits in existing hierarchies [8], [35]. Godin [24] developed FCA algorithms to infer a non-redundant form for implementation and interface hierarchies and carried out experiments on several Smalltalk applications. For dealing with UML models, Relational Concept Analysis, an extension of FCA, takes the relations into account [27]. Other approaches analyze class hierarchies using access or usage information. [49], [51] analyze the usage of the hierarchy by a set of client programs. Mining aspects has been considered in the context of FCA [10].

Current Issues — Factorization is by nature a combinatorial process. Recent studies [21] show that Relational Concept Analysis, applied to rich UML descriptions including references between concepts, produces a huge number of artefacts which is quite impossible to analyze by hand. Execution time can become a problem for large size software, but the actual difficulty is the result size.

Other modular construct discovery: Clustering approaches are, by far, the preferred algorithmic approach to the problem. They have been proposed to identify modules in applications that are not specifically object-oriented (e.g. [29], [30], [38], [43]).

Finding components in object-oriented software is proposed in [12]. Simulated annealing is used to gather classes into components by optimizing metrics measuring cohesion and coupling. The problem has similarity with package mining.

It is a well-known practice to layer applications with bottom layers being more stable that top layers [41]. Until now few works have been done in practice to identify layers: Mudpie [55] is a first cut at identifying cycles between packages as well as package groups potentially representing layers. DSM (dependency structure matrix) [53], [46] are adapted for such a task but there is a lack of detail information. From the side of remodularization algorithms, a lot of them were defined for procedural languages [31]. However object-oriented programming languages bring some specific problems linked with late-binding and the fact that a package does not have to be systematically cohesive since it can be an extension of another one [18], [57].
C. Techniques for module assessment

Software Metrics: Re-modularization of software systems is geared toward producing highly cohesive and loosely coupled modules. Many different cohesion/coupling metrics were proposed (including a study by [2]). In the more specific case of object-oriented programming, assessing cohesion and coupling has been the focus of several metrics. However, their success is rather mitigated as the number of critics raised. For example, LCOM [13] has been highly criticized [4]. Other approaches have been proposed such as RFC and CBO [13] to assess coupling between classes. However, many other metrics have not been the subject of careful analysis such as Data Abstraction Coupling (DAC) and Message Passing Coupling (MPC) [5], or some metrics are not clearly specified (MCX, CCO, CCP, CRE) [36]. New cohesion measures were proposed [40], [45] taking class usage into account. The Cohesion/Coupling dogma, however, started to receive critics in recent times [1], [47]. People argue that software engineers do not base clustering on this criterion but rather use more semantical approaches.

Software Visualization: There is a significant effort to create efficient software visualizations to support the understanding and analyses of applications [32], [37], [39], [56]. Lanza and Ducasse worked on system level understanding combining metrics and visualization [33] and class understanding support [16].

Current Issues – These approaches are often not customized for object-oriented applications. Existing solutions propose modules at a very low level of abstraction that do not reduce enough the size of the system comprehension problem. Solutions that may offer larger (potentially more abstract) modules, result in modules that have no meaning for the software engineers.

Complementary Remodularization Algorithms: There is a need for a global remodularization infrastructure in terms of analyses (algorithms, information presentation, metrics) that can take into account the diversity of the abstractions in presence (different module semantics, different abstractions and relationships including different levels, functions, classes, packages, etc.).

Complexity and approximation: In the point of view of graph theory, the central problem seems to be close to the classic $k$-cuts problem [23], [25]. Nevertheless, we must add several new criteria, like the quality of the proposed solution. The first step of the research will be the characterization of an optimization criterion. We also think useful to study and analyse the sensibility of the problem with respect to the operations: add/delete of vertices/edges. It is a major challenge to be able to propose robust approximations.

Scalability: Computational complexity of algorithms has a limit, already known for Formal Concept Analysis (FCA/RCA) and foreseeable for exact methods. Checking the scalability of these algorithms is thus an additional challenge. Another issue is the combinatorial explosion which may occur in the remodularization results. Because of the size of current applications, presenting these results to the engineers and guiding them to take a decision is an additional challenge.

Reengineer inputs and quality of the solution: Engineers should drive the remodularization. Fully automated approaches are applicable only to a very limited context. In reality, external constraints have to be specified and taken into account by the remodularization algorithms (such as the inclusion of a class in a specific package). Software engineers should guide the process possibly confronted to different solutions and their relative impacts. Often favoring minimum impact on existing code has to be considered. Finally the quality of the resulting modularizations has to be taken into account.

As a conclusion: The problems of software evolution are many and varied, in this proposal we plan to consider one of these problems: software remodularization. We think urgent to drive such a complete study of the problem, both “vertically” by studying all the aspects of the modularization problem (modeling of the software, modularization quality metrics, modularization algorithms, presentation of the results), and, “horizontally”, by considering different modularization approaches. The solution will not apply one single method, but a combination of various skills in different research domains. Such a research would also be guided by platforms for testing ideas on real-world applications.


