An Approach for Empirical Evaluation of Model-Driven Engineering in Multiple Dimensions

Parastoo Mohagheghi (SINTEF)

Problem

- Methodologies, tools or techniques must be evalauted:
  - From multiple points of views of different stakeholders;
  - By using multiple approaches and multiple data sources;
  - In multiple contexts;
- Two motivations:
  - Usability in one specific context;
  - Generalization to other contexts (projects or domains).
- There have so far been few empirical studies of MDE.
  - A systematic review performed by us in 2007 managed only to detect 25 papers on applying MDE in industry published between years 2000 and 2007 and most of these were related to small-scale studies.

This paper's goal

- Presents the evaluation method applied in the MODELPLEX project.
- MODELPLEX (MODelling solution for comPLEX software systems) ran in 2006-2010.
- Large IP, four industrial cases in four domains with complex cases to develop.
  - SAP, Thales, Telefónica and WesternGeco
- We also present some results that are not included in the paper.

Requirements of the evaluation process

- MDE is a generic approach that is applied in multiple ways in organizations:
  - The evaluation method should therefore be context-dependent.
- Industrial partners may have problems in revealing the details of their development processes and results:
  - The evaluation method should allow collecting some general results.
- EU wants to evaluate the exploitation of project results in industry and intentions for future usage as a measure of project success.

Evaluation approach

- The ”evaluation plan” has three distinct parts:
  1. *Research questions* that were defined by each industrial partner and reflect their goals with applying MDE. **Addresses context-dependence.**
  2. *Technology Acceptance Model* (TAM) that is used for collecting the developers’ perceptions regarding MDE and tools, and intensions for future usage. **Addresses exploitation.**
  3. Evaluating tools by assigning scores by users in order to collect some **general feedback** and measure the success of individual tools in meeting users’ requirements.

First:

RESEARCH QUESTIONS

Identifying evaluation criteria in reserach questions

- A bottom-up process that starts from analyzing model-driven engineering techniques (Practices) and what do they provide (Promises);
- Four practices are identified:
  1. *Models everywhere:*
     + improved communication between stakeholders and improved software quality by using models for early analysis and testing
     - modelling has a cost and modelling tools must be integrated with other tools
  2. *Multiple abstraction levels and separation of concerns in models:*
     + improved communication, portability of solutions
     - keeping models consistent with one another

Continue on MDE practices

- *Two other:*

3. *Generating Artefacts from Models:*

   + supports separation of concerns and adding details later; not by manual work but by applying transformations.

   + less manual work, consistency and traceability between artefacts

   -The cost relies in developing transformations.

4. *Metamodeling:* allows definition of new modelling languages or extending the existing ones:

   + Sharing the same language between domain and IT experts and narrowing the gap between them, and involving domain experts in all stages;

   + exchanging models between tools

   - defining metamodels and supporting tools requires high initial investment and needs language and tool expertise.

Examples of evaluation criteria

Subject of evaluation	Related to MDE practice	Evaluation criteria in research questions	Data collection
Modelling framework based on a metamodel for defining architectural models in different views	Metamodeling, Multiple concerns,	Does the framework support separation of concerns?	Apply on appropriate scenarios.
		Are the views expressive enough?	Check for criteria: including necessary concepts, modelling dependencies between views,
Using models for performance simulation	Models everywhere, Metamodeling	Are the performance results comparable with the actual performance of the system within +/-25%?	Exploratory case study and comparison
		Is it possible to integrate the performance modelling with testing tools?	Interoperability via XMI should be evaluated in a scenario.
		How complex is performance modelling (ease of learning)?	The concepts should be learnt in less than 4 hours in an experiment.
DSL for network modelling	Metamodeling, Generating artefacts from models	How efficient is the model transformation process?	Time taken in writing transformations in a case compared between tools.
		How readable is the generated code?	Compare with manually written code. Do the artefacts need any post-processing?

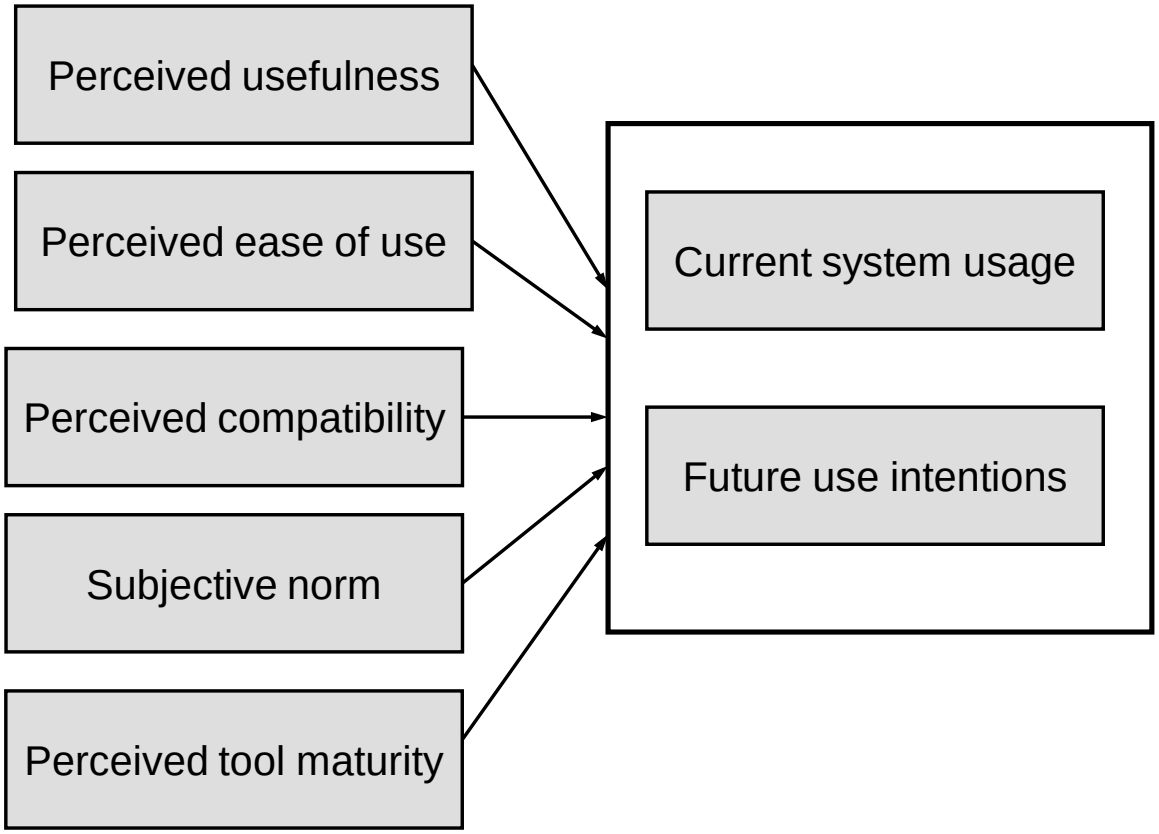
Results: Where does MDE help?

- *Simulation and Execution based on models eveywhere and generation*:
- *Abstraction*:
  - Partners have all taken advantage of abstraction in developing their complex systems.
- *Communication based on metamodeling*:
  - Several cases in MODELPLEX involved developing DSLs or UML profiles for improving communication.

Second:

TECHNOLOGY ACCEPTANCE MODEL

The Applied TAM Model



Data collection

- Semi-structured interviews in the beginning of the project to map the state of the practice before project.
- An on-line survey at the end of the project to collect data regarding perceptions, the state of actual usage and future intensions of use.
- Examples of questions:
  - I would like to use the MDE approach in the future for my work.
  - People who are important to me think I should use the MDE approach.
- five scales; from “strongly disagree” to “strongly agree”.
- We analysed the results qualitatively and could not run statistical tests.

State of the practice before MODELPLEX

- All the companies had experience with modelling for the purpose of analysis and design while practices of MDE were applied in varying degrees:
  - Most companies had ad-hoc code generators or generators integrated in their development environment for generating skeletons and stubs, user interfaces and some test cases.
  - Companies are familiar with Model-to-Model (M2M) transformation
- All companies have experience with Eclipse as a development tool for Java programming.
- In many areas of MODELPLEX research, companies had little or no experience; such as Model repository and Model-based V&V.

Some results

- MDE is generally perceived as useful for solving the problems of users while it is not perceived as easy to learn.
- Regarding tool maturity, performance and functionality of tools were generally perceived as satisfactory, while scalability to large projects is a concern.
- The participants generally agreed with the statements that they would like to use MDE in their future work.

Third:

ASSIGNING SCORES TO THE TOOLS

Data collection

- We asked participants to give scores from 1 to 5 (where 5 are best) to the tools in five dimensions: functionality, ease of use, compatibility, performance and reliability, and total impression.
- We covered tools developed in MODELPLEX and some external tools and technologies that were widely used such as Eclipse.

Some results

Reuseware composition tool used on an experimental basis

	Functionality	Ease of use	Compatibility	Performance & Reliability	Total impression	Intended Future use
Telefónica	5	4,5	5	4,5	4,5	Yes
Thales	4	3	3	3	3	Probably yes

Tools should be improved regarding multi-user support, versioning of models, conflict detection and resolution, diff/merge possibilities, and modelling at different levels of abstractions. All of these features are required for using tools in large and complex projects.

Finally:

CONCLUSIONS

Advantages and limitations

- The evaluation method provides several sources of data to analyse and combine.
- Some threats to the validity of the results are:
  - Construct validity is concerned with “right metrics”. ”. We mean that applying the Practice-Promise approach improves the construct validity.
  - A methodology such as MDE impacts different constructs. The approach proposed here takes advantage of multiple criteria and several data sources and thus reduces this risk.
  - Several research questions are answered by experts and based on their experience and expertise. Involving different people in the evaluation can reduce the threat of subjective judgements.
  - Performing comparative case studies is costly but useful.
  - The expectations of companies participating in a research project may impact the results.

Relations to other work

- Other approaches such as GQM start with goals and not an analysis of the methodology and what we can expect.
- Popular metrics such as productivity and defect rate are not always relevant or possible to collect.
- Earlier studies do not have a systematic approach to select evaluation criteria.
- Models such as TAM are useful in collecting perceptions but they should be combined with more specific measures in order to explain the results.

Innovations?

- Two types of feedback are provided: The *observed values* based on research questions and the *perceived values* based on the survey, which can explain one another. We have related metrics to TAM dimensions.
- Several validity threats are reduced by taking advantage of multiple data sources and multiple evaluation criteria.
- One of the major requirements of EU projects is to measure project success by infusion of the results in industry. We have mapped the state of the practice before the project and intensions for future usage in order to evaluate the project success.

Evaluating usefulness

Criteria	Definition
Architecture model quality	Whether the architecture model satisfies criteria such as support for separation of concerns and integration of several viewpoints.
Design quality	The solution improves the quality of design by identifying poor design.
Quality of generated artefacts	The quality of code, documentation etc. that are generated from models is acceptable (understandable, compliant with coding standards etc.).
Solution suitability	The solution can solve the problem in hand.
Generation ratio	The number or size of the generated elements divided by the total number of elements or size. The question is whether the saving by generation can compensate for modelling effort.

Evaluating compatibility of Solutions

Criteria	Definition
Cost of adoption	Effort required for setting up a tool and customizing it.
Integration with other solutions	The degree a solution can be integrated with other practices or tools.
Standards compliance	A solution must conform to selected standards.

Thank you and

Questions?