DOC2MODEL

1 Introduction

This document is a docuware for creation review and describes the doc2model project.

2 Aim

The Doc2Model (Document to Model) framework is a proposed open source component under the Eclipse Modeling Framework Technology project for parsing structured documents (e.g., xlsx, docx, odt, odf...) to produce EMF models. It is in the Project Proposal Phase (as defined in the Eclipse Development Process document) and is written to declare its intent and scope. This proposal is written to solicit additional participation and input from the Eclipse community. You are invited to comment on and/or join the project. Please send all feedback to the Eclipse Modeling Framework Technology (http://www.eclipse.org/newsportal/thread.php?group=eclipse.technology.emf) newsgroup with [doc2model] as a prefix of the subject line.

3 Background

The most widely used tools in many organizations continue to be text processors and spreadsheets. Often these documents describe business data that are important to manipulate in other contexts. Examples of data contained in such documents include the following:

- Requirements
- Tests
- CRC cards
- Structure definitions
- Documentation generation

Because these kinds of tools often produce plain text documents, it's typically quite complex and time-consuming to develop a specific parser able to produce output more amenable to further manipulation.

Currently some organizations are investing effort to publish specifications of open source file formats, for example Office XML (e.g., docx, xlsx...) and Open Document (e.g., odt, odf...) to facilitate widespread adoption and easier consumption.

In fact, most of the business documents are organized with data defined in a common way, (top down for example for text documents) using text style, regular expressions, and column numbering. As such, it's possible to support a generic solution for parsing those documents and transforming the business data into EMF models, using XML parsing and EMF's reflective capabilities.

4 Scope

This project will provide an extensible framework for producing EMF model instances from plain text and structured documents.

Transforming a business document into an EMF model will facilitate more opportunities to exploit the business data contained in such a document. In some cases documents represent the specification of a system. Instead of retyping information to produce the corresponding model it will be possible to generate it.

Doc2Model can be used to, for example, to import requirements from text files and transforming them into SysML requirements models.

The documents file formats which will be managed by Doc2Model include

- Open source formats as docx, xlsx, odt, odf;
- Common formats as csv;
- And formats desired by the eclipse community.

The Doc2Model API will provide extension mechanism to allow users to add custom parsers for specific tools. These Parsers could be contributed to Doc2model component if the license is compatible with EPL.

5 Description

The target model type is specified using a configuration model which describes how the data is identified during the parsing. This configuration is a map indicating what the generator does when a matching rule is applied. Matching rules make use of regular expressions, special styles, columns (spreadsheet), and tags. Transformation proceeds as follows:

1. Read the matching configuration.

2. Analyse (parse) the input document to identify matching data base on the rules.

3. Produce an output EMF model instance from the data recognized into the input document. Cross references between the data is supported and additional data can be injected.

6 Sample

- input user document

1 Presentation

2 Requirements

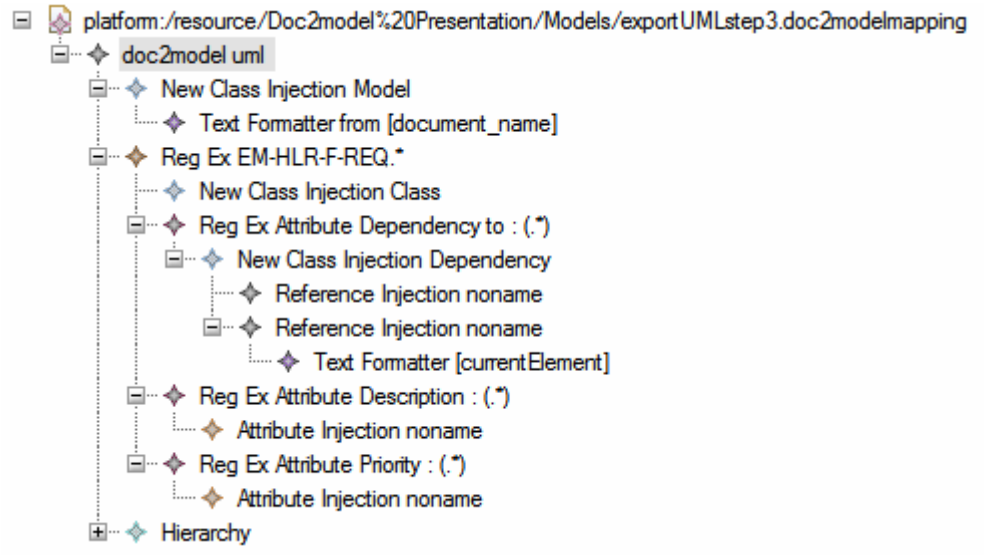
2.1 Level A

EM-HLR-F-REQ-001
Name : Equipment state
Description : Equipments regularly send signals (a frame) to give their state. Signals differ according to the equipment category.
Priority : Mandatory
EM-HLR-F-REQ-002
Name : Case of failure
Description : In case of failure assumption (particular signal), advanced analysis should be performed and correlation should be done with other existing information to check the effective failure.
Priority : Mandatory
EM-HLR-F-REQ-003
Name : Handle failure
Description : If failure is confirmed, a procedure should be presented to the maintenance operator to handle the failure.
Priority : Mandatory

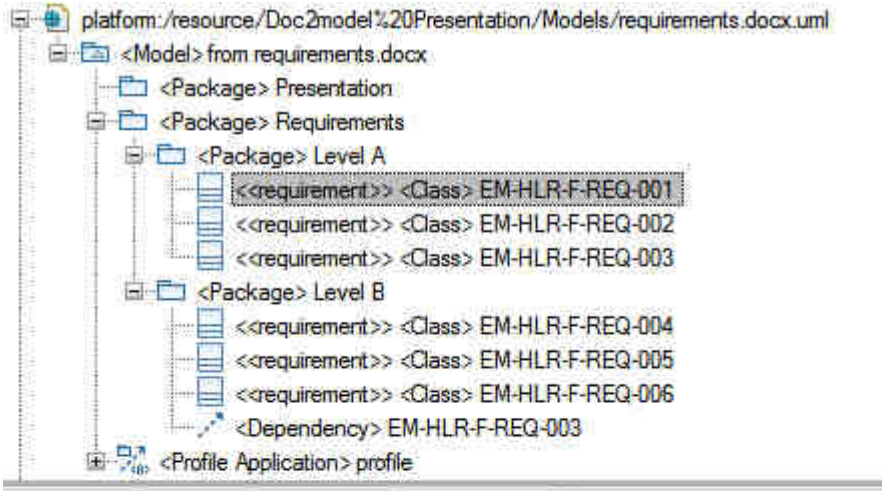
2.2 Level B

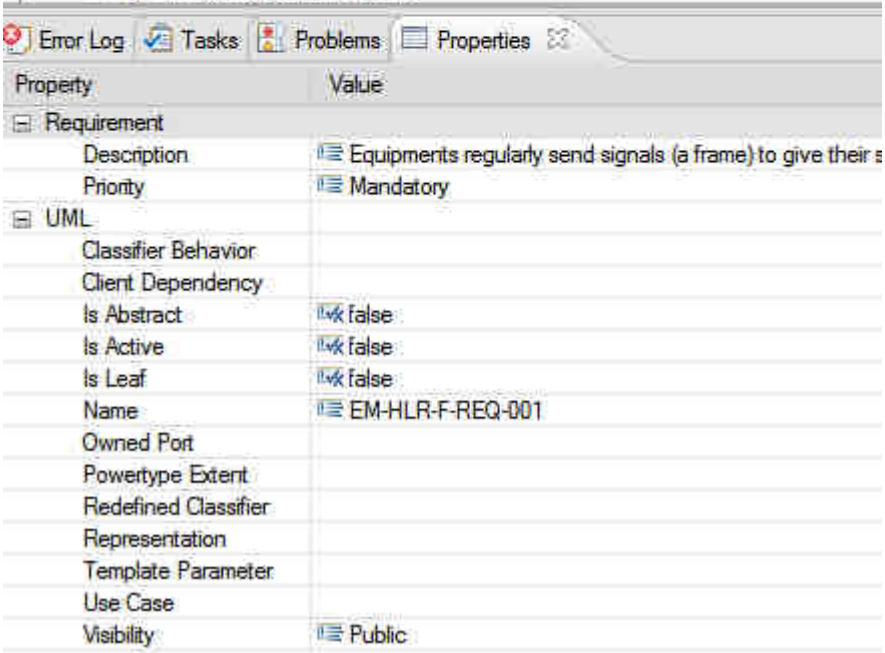
EM-HLR-F-REQ-004
Name : Procedure
Description : The procedure differs according to the equipment type.
Priority : Mandatory
EM-HLR-F-REQ-005
Name : Permissions
Description : The target system should allow administrating the equipments and their categories.
Priority : Mandatory
Refine : EM-HLR-F-REQ-002
EM-HLR-F-REQ-006
Name : Failures statistics
Description : Statistics about failures and incident resolutions should be activated if needed and should be exported as reports if requested by the chief operator.
Dependency to : EM-HLR-F-REQ-003
Priority : Mandatory

- doc2model mapping



- result after execution





7 Relationship with Other Eclipse Projects/Components

- Doc2Model will be built on top of EMF.
- Doc2Model will exploit EMF Compare to obtain differences between models.

8 Third party libraries

- no third party librairies, standard java parsing.

9 Code Contributions

- Topcased is offering doc2model as an initial codebase (see http://gforge.enseeiht.fr/projects/doc2model).
    - The code is hosted at : http://gforge.enseeiht.fr/projects/doc2model
    - There is no existing domain names associated with this code.
    - There is no trademarks applied to the project name.
    - Doc2model was hosted under EPL license.
    - present contributors and committers:
        - Benjamin Marconato: benjamin.marconato@atosorigin.com
        - Thibault Landré: thibault.landre@atosorigin.com
        - Emilien Perico: emilien.perico@atosorigin.com
        - David Ribeiro Campelo: david.ribeirocampelo@atosorigin.com
        - Tristan Faure: tristan.faure@atosorigin.com
- a flash demo of current version is available here, http://www.4shared.com/file/123030374/bd993fe4/doc2model.html

10 Organization

10.1 Committers

- Tristan Faure (Atos Origin), tristan.faure@atosorigin.com

Tristan Faure is a software engineer at Atos Origin. Working for 3 years with Eclipse Technology and modelling (EMF, UML...). He is also an EMF, OCL and ATL consultant.

Commiter and technical manager for TOPCASED open source project, he makes training about TOPCASED platform for international Atos Origin teams.

- Emilien Perico (Atos Origin), emilien.perico@atosorigin.com

Emilien Perico is a software engineer at Atos Origin, working for 3 years with Eclipse platform and modeling technologies such as UML2, GEF and GMF and model transformation tools.

He is a committer on TOPCASED project, has been in charge of the platform build and packaging. He also worked on projects about tests generation and document generation. He is now a committer on the Eclipse MDT Papyrus project.

- Werner Keil, werner.keil@gmx.net

*Werner Keil is currently **Eclipse** RCP Develper and Consultant at mobilkom / **Vodafone** after having worked for governments or Fortune 500 companies like Daimler, Nokia, BEA/Oracle, GE, Shell, Commerzbank, Capita, Credit Suisse or Legal & General.*

He has worked for more than 20 years as project manager, software architect, analyst and consultant on leading-edge technologies for Banking, Insurance, Telco/Mobile, Media and Public sector. Among earlier clients are Sony Music where Werner designed and implemented micro-format based tags for their online music portal.

He develops enterprise systems using Java, JEE, Oracle, IBM or Microsoft, does Web design and development using Adobe, Ajax/JavaScript or dynamic languages like Ruby, PHP, etc. Besides work for major companies he runs his own creative, talent and consulting agency Creative Arts & Technologies. In his spare time, Werner runs and supports open-source projects, writes song lyrics, novels, screenplays and technical articles.

*He is committing member at Eclipse Foundation, **Babel Language Champion** (German) and active member of the **Java Community Process**, including his role as JSR-275 Spec Lead, JavaEE 6 EG and **Executive Committee Member** (SE/EE).*

10.2 Mentors

- Ed Merks, ed.merks@gmail.com
- Cedric Brun, cedric.brun@obeo.fr

10.3 Interested Parties

- Atos Origin, http://www.atosorigin.com/
- Obeo, http://www.obeo.fr/