

O sistema de acesso a bases de dados pelo telefone (denominado “Comunica”) é um projeto de pesquisa e desenvolvimento que está em andamento no Instituto de Informática da UFRGS, sendo fruto de uma parceria entre as empresas Conexum (encubada no Centro de Empreendimentos), DFL e Intext Mining. O comunica é um sistema com reconhecimento e síntese de voz que permite ao usuário consultar informações em base de dados. O sistema está dividido em vários módulos, e um deles é o ETL (Extract-Transform-Load), encarregado de extrair as informações de uma base de dados estruturada, transformá-las para um formato mais adequado e carregá-las em outro banco de dados. O módulo é importante porque o Comunica não pode ser utilizado diretamente em uma base de dados de produção, tradicional. Ele necessita de informações estruturadas em uma forma específica, organizadas com base em uma ontologia de domínio, que é desenvolvida caso-a-caso, de acordo com o banco de dados alvo. Logo, o módulo ETL faz a importação dos dados da base operacional do cliente em uma base de dados estruturada do projeto Comunica. Para tanto, códigos (scripts) ETL foram desenvolvidos utilizando a linguagem PHP. Eles não só traduzem mas também validam os dados antes de serem inseridos no banco do projeto. Caso os dados não estejam corretos por algum motivo, o erro é registrado em um arquivo de log e enviado por e-mail. Tais scripts foram automatizados através de rotinas Cron. Os scripts gerados foram adaptados de outros já existentes, criados por Fabio Moreira da Silva, Thiago Cardoso Amador e Daniel Nehme Muller. A principal tarefa foi automatizá-los, incluindo o registro do status em arquivos de log e validações. Atividades futuras incluem a generalização do ETL para que ele possa facilmente ser implantado em mais clientes, atualmente ele está executando apenas em um banco de dados cliente. Além disso faz-se necessário aprimorar as técnicas de recuperação de erros humanos, tais como erros de digitação de meta-dados que são frequentes e no estado atual requerem correção manual.