166 **EXTRAÇÃO DE ESQUEMAS DE DADOS DE DOCUMENTOS SEMI-ESTRUTURADOS.** *Diego Menin* (1), *Renata de Matos Galante* (2) *Sandra Rovena Frigeri* (3) (1) Bolsista de Iniciação Científica Fapergs (2) Orientadora (3) Co-orientadora.

O projeto KDD-XML (Análise de Informação em Dados Estruturados com XML) tem por objetivo propor um arquitetura de integração de bases de dados através do uso do padrão XML para disponibilizar informações na web. Esse ambiente possuirá um conjunto de ferramentas, onde a mais visível ao usuário final será a que realizará consultas convencionais e consultas de análise de dados, as quais serão realizadas através da implementação de recursos para descoberta de conhecimento em bases de dados (DCBD). Nesta parte do projeto estão sendo implementadas ferramentas que permitirão extrair a estrutura de um documento html, gerando esquemas de dados em XML-Schema e documentos XML, os quais poderão novamente ser transformados em documentos html conforme padrões de exibição pré-especificados em documentos XSL. O objetivo principal da extração será identificar no documento html algumas estruturas pré-definidas como elementos de informação, os quais orientarão a construção do XML-Schema. A partir do esquema de dados de uma classe de documentos, será possível a extração de informações de diversos documentos que contenham os mesmos elementos de informação, mas que possuem diferentes formas de apresentação dessas informações. Esses elementos poderão ser armazenados em bases de dados e posteriormente acessados por mecanismos de consulta. O extrator foi organizado nos seguintes módulos: analise de um documento XML, analisando se este é bem formado e gerando o seu esquema de dados em XML-schema; extração de um documento XML a partir de um documento html, utilizando com referência o esquema de dados da classe de documentos html; estruturação de documentos XSL para definição das estruturas para visualização de documentos XML; e geração de documentos html, utilizando documentos XML e estruturas de visualização XSL.