XML to Graph Mapping Tool

The ISEBEL project aims to build an international search engine that is able to harvest data from folktale databases. The initial project concentrates on belief legends [1] found in the three well known digital collections by Evald Tang Kristensen from Denmark (etkspace), Richard Wossidlo from Mecklenburg (wossidia) and several collectors and narrators from the Netherlands (verhaalenbank). Part of the project is on data and graph mining [2,3] for frequent patterns. Therefore, the XML story data is harvested using OAI-PMH and transformed into graph data.

The WossiDiA system [4] itselfs as one of the databases harvested by ISEBEL uses typed, directed hypergraphs [5] for representing the collections by Richard Wossidlo. The content encompasses field notes, correspondences with scholars, contributors and informants as well as references to published work on the everyday life in the country Mecklenburg from late 19th century to the 30s of 20th century.

To support the researchers in analyzing, browsing and visualizing certain aspects in the collected data using the graph paradigm and algorithms developed not only restricted to the WossiDiA graph data, the master thesis should aim at transforming all collected XML data into property graph data [6].

Therefore, a general methodology has to be developed for transforming XML data into graph data. The transformation should be guided by user defined rules which describe a mapping from XML Schema concepts to Property Graph Model concepts [7, 8]. A rule specification languages has to be designed. By that, the ethnologist and researcher should be able to define rules which select elements, attributes, and content from XML files and generate nodes, labels, edges, and properties of a graph from that selection. Finally, a tool has to be developed which automatically generates the property graph data based on the rule-set given by the user. It functionality has to be demonstrated by sample scenarios from the ISEBEL project, e.g. the witch and werewolf hunter scenario looking for the gender influence on witch or werewolf stories.

Road map

- Research on, analysis of and summing up the XML data and Property Graph Model
- Presenting the state-of-the-art in XML and graph transformation concepts, techniques and tools
- Requirement analysis of graph visualization scenarios of the ISEBEL project
- Defining a transformation rule language
- Designing a software tool for rule-based transformation
- Prototype implementation and quality based evaluation using ISEBEL sample scenarios

Character

State-of-the-art analysis, concept design, prototype implementation

Prerequisites and technologies

Graph models and databases, programming language Java or Python

Advisor

Holger Meyer

References

1. Usó-Doménech, J.L. & Nescolarde-Selva, *What are Belief Systems?*. J. Found Sci (2016) 21: 147.
2. Charu C. Aggarwal, Haixun Wang: Managing and Mining Graph Data. Advances in Database Systems 40, Springer 2010, ISBN 978-1-4419-6044-3
3. Diane J. Cook, Lawrence B. Holder (eds), Mining Graph Data. Wiley, Hoboken, New Jersey, 2006
4. Holger Meyer, Alf-Christian Schering and Christoph Schmitt, *WossiDiA --- The Digital Wossidlo Archive*, in: Holger Meyer, Christoph Schmitt, Thomas Jansen and Alf-Christian Schering (Hrsg.), Corpora ethnographica online --- Strategien der Digitalisierung kultureller Archive und ihrer Präsentation im Internet, Volume 5 of Rostocker Beiträge zur Volkskunde und Kulturgeschichte, Waxmann, 2014, 61–84.
5. Meyer, Holger, Alf-Christian Schering, and Andreas Heuer. "The Hydra. PowerGraph System." *Datenbank-Spektrum* (2017): 1-17.
6. Angela Bonifati, George Fletcher, Hannes Voigt, Nikolay Yakovets: *Querying Graphs*. Morgan & Claypool, Synthesis Lectures on Data Management, 2018.
7. Genoveva Vargas-Solar, José-Luis Zechinelli-Martini, Javier A. Espinosa-Oviedo: *Enacting Data Science Pipelines for Exploring Graphs: From Libraries to Studios*. ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium, 271-280, 2020.
8. Dominik Tomaszuk, Renzo AnglesŁukasz Szeremeta, Karol Litman, and Diego Cisterna: *Serialization for Property Graphs. Beyond Databases, Architectures and Structures. Paving the Road to Smart Data Processing and Analysis.* BDAS 2019: Beyond Databases, Architectures and Structures. Paving the Road to Smart Data Processing and Analysis, 57-69, 2019.