Design Concept

4.1 Problem Definition

We are presented with the ISEBEL project, a digital archive of stories from belief legends found in three well known digital collections by Evald Tang Kristensen from Denmark (etkspace), Richard Wossidlo from Mechlenburg (wossidia) and several collectors and narrators from the Netherlands (verhaalenbank). These databases are made up of stories originating from different source. Stories are composed by different authors and spread across many papers in the database, stories in these papers have facts and contents which are related and also an author may contribute related ideas to content of different papers. Thus, these databases are stock with varieties of stories, with some stories so inter-related such that comprehensive information cannot be found in a single document, stories have to be compiled from documents all across the archive. Facts and contents of the stories may be related either by the stories itself or by author and co-authorship.

Therefore, we are faced with the challenge of modeling and implementing a framework to present the users of ISEBEL search system with a more relevant search result that effectively gives results of stories, related stories, possible stories that will be interconnected in the feature and a way to visualize stories inter-relatedness through the use of data mining techniques.

4.2 Data Extraction process

One of the key steps towards a successful data mining is the availability of data. In this research, the XML story data to be used is harvested from the three databases (wossidia, verhaalenbank and etkspace) using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). The OAI-PMH defines an open interface for exchange of metadata. It has an architectural model that allows **data providers** make metadata available through a well-defined protocol. The metadata exposed by the **data provider** allows the **service providers** to harvest it and then aggregate it, post-process it, and refine it with the goal of developing services that add value (see figure 4.1).

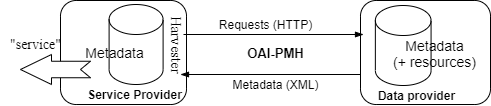


Figure 4.1: OAI-PMH Architecture

For this project, communication takes place between the three databases (wossidia, verhaalenbank and etkspace) as the data provider and the ISEBEL archive as the service provider. See the pictorial representation of the communication topology in figure 4.2.

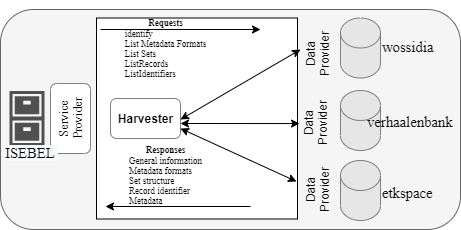
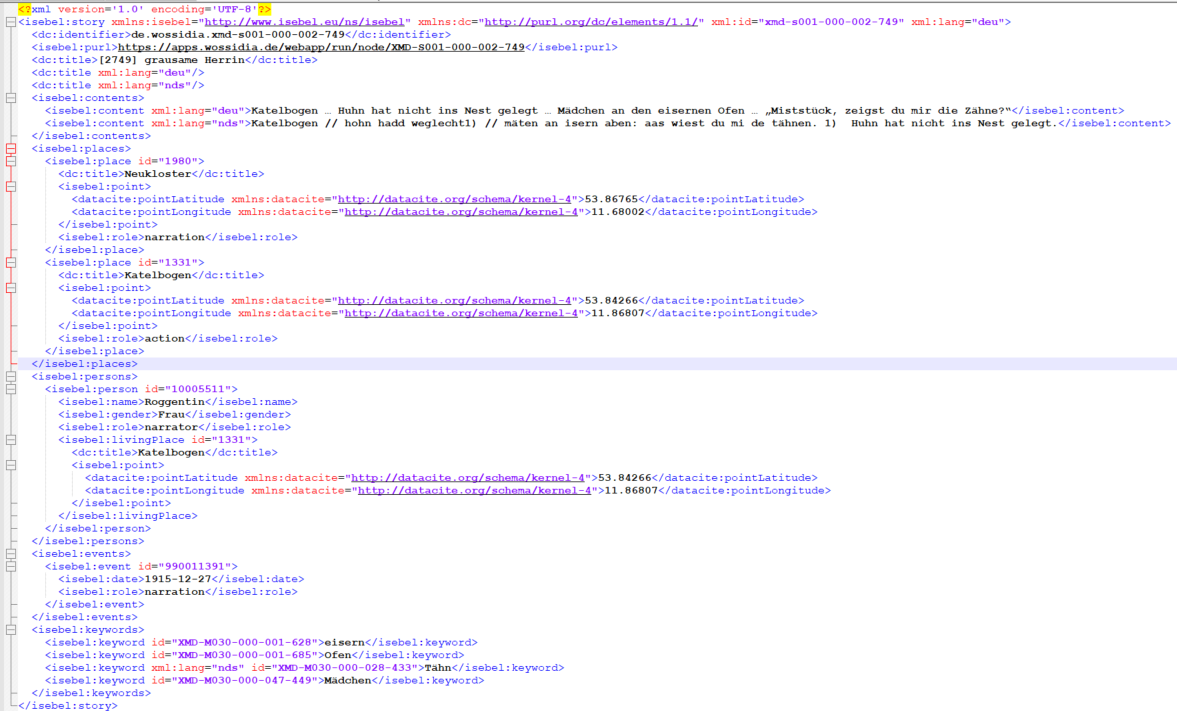


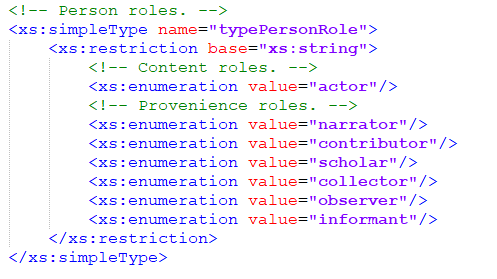
Figure 4.2: OAI-PMH Architecture for ISEBEL

In simple terms, the OAI-PMH has been implemented in our system which allows access to metadata in XML format made available from different sources and is used for further analysis in this project.

4.3 Dataset Analysis



In order to make all the sources of data that are harvested and made available in the ISEBEL archive to be realizable as a graph data, a general schema has been defined which sets the minimal requirements for the structure of the ISEBEL XML story document. Hence, all sources must use this schema to provide their stories. The schema defines all the elements and attributes that are valid in the XML documents. It also specifies tags that are allowed within another tag. A snippet of a portion of the XML document for Person roles schema can be shown in code 4.1. This schema describes an enumeration of possible role values that a person authoring a story can have. What this means is that, outside the possible enumerated roles of actor, narrator, contributor, scholar, collector, observer, and informant, the schema cannot accept any other value from a source for the person role. Likewise other schema in the document, though there are still some schemas that does not have restricted values.



Code 4.1

Of more importance to this thesis is how schemas can be related across document to form an interconnection which helps to relate stories across documents that eventually forms the nodes and edges of graph data.

In order to understand and make a more informed decision about the XML story data harvested from the ISEBEL archive, these data have been further converted to a comma separated value (CSV) file so it can be processed as a graph data.

4.3.1 WossiDiA dataset

The WossiDiA dataset is a dataset extracted from WossiDiA digital archive information system. The archive contains more than 2.5 million digital presentation of Richard Wossidlo’s folklore, an ethnologist and ethnographer who in his study gathered stories related to the ancient customs originating from Mecklenburg between 1883 and 1939.

WossiDiA architecture is based on a multidimensional graph database called **hypergraph database,** which makes its highly complex network structure collection visible and enables innovative navigation, search scenarios and the visualization of results.

The archive is rich with stories that narrates people’s cultural heritage from different authors in different places and time.

For the purpose of this thesis,