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).

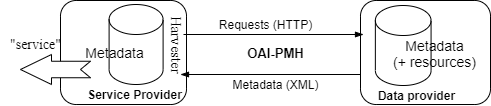


Figure 4.1: OAI-PMH Architecture

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.

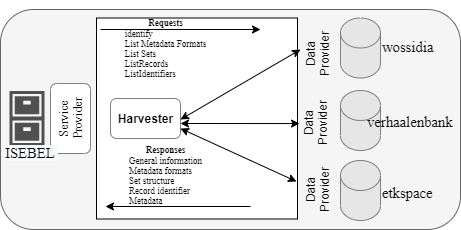
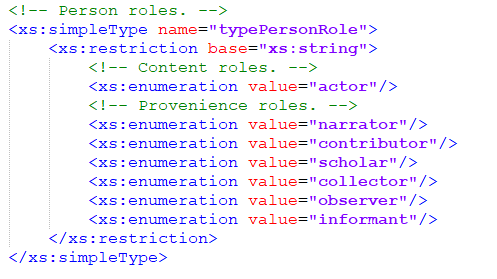


Figure 4.2: OAI-PMH Architecture for ISEBEL

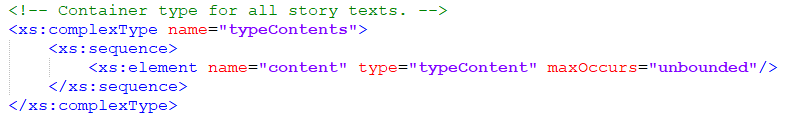
In simple terms, the OAI-PMH has been implemented in our system which allows access to metadata in XML format made available from different sources and is used for further analysis in this project.

4.3 Dataset Analysis

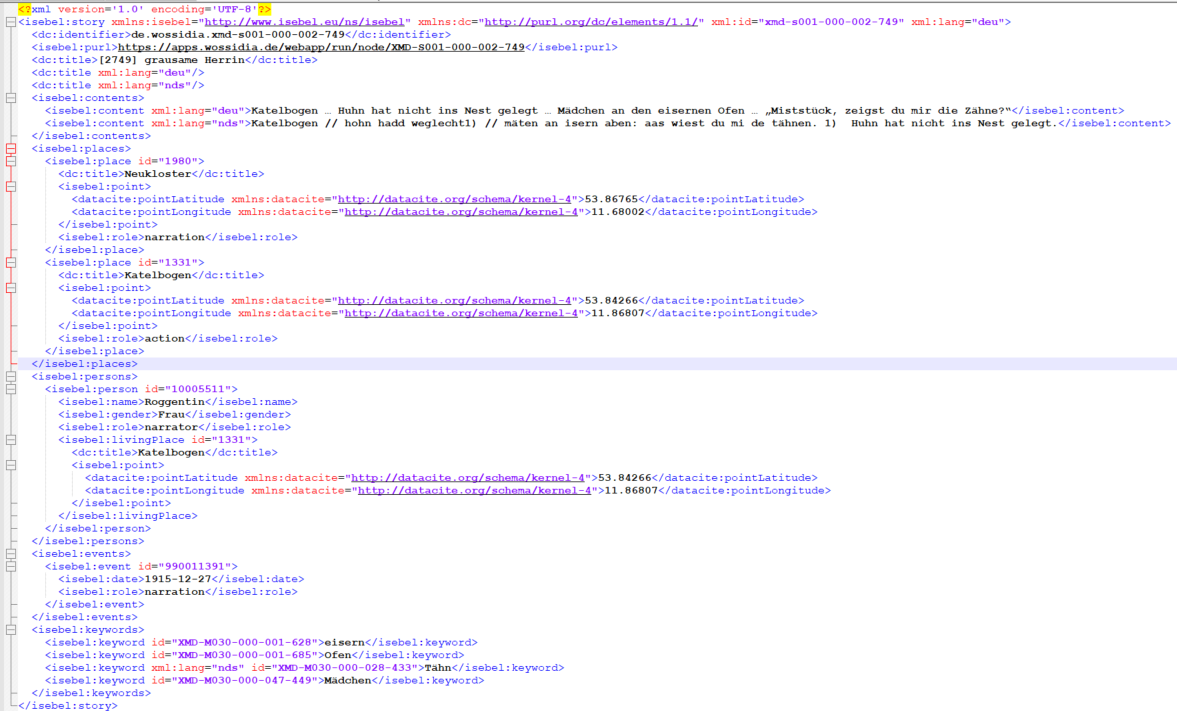
In order to make all the sources of data that are harvested and made available in the ISEBEL archive to be realizable as a graph data, a general schema has been defined which sets the minimal requirements for the structure of the ISEBEL XML story document. Hence, all sources must use this schema to provide 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 element can be shown in code 4.1. This element 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 element cannot accept any other value from a source for the person role. Likewise other elements in the document, though there are still some elements that does not have restricted values, such as the element for story texts in code 4.2. Code 4.3 show an extract of an XML story document from wossidia where all the elements defined within story root element has been represented to form a complete XML story document.



Code 4.1



Code 4.2



Code 4.3

Of more importance to this thesis is how elements in a schema 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. A close look at wossidia XML document shows that some tags such as person, event and place have an ID which can be used to identify them. It means that for two stores authored by the same person or in same place or during same event but represented in different document, there exist a possibility of the two stories having same person ID or event ID or place ID or a combination of two or maybe even the three. Thus, an edge has been created between the two-story documents and that is how graph is formed using different XML story documents, see illustration in figure 4.3.

Also, for elements without an identifier such as the content tag, edges can still be formed using such elements by mining related keywords of texts placed in-between the tags using a method of natural language processing. However, employing the natural language processing method to obtain these fine grain details maybe above the scope of this study, thus we will not discuss it further.



Figure 4.3: XML Story document relationship

From visual analysis of this XML document, two conclusions can be drawn:

* XML story documents/ story schemas are likely to represent graph nodes
* Elements attributes like person id, event id, place id and content keywords are likely candidates for graph edges.

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 easier 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[1]. 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, the dataset has been extracted from the archive in XML format and converted to CSV files. Two CSV files are presented one containing the node details and the other containing the edge list. Table 4.1 and 4.2 shows the first ten rows of the node and edge dataset respectively. To understand the dataset better, a comparison can be made between the two table and a gephi file extract made with the combination of the two tables, it can be observed that the node table has an ID, label and type which represents the nodes id, name and type of interconnection relationship with other nodes respectively. Matching this observation with the edge table, we notice that the edge table has a source and target which shows the nodes that are connected to each other and a label which is the attribute that connects the two nodes. In other to understand this node-edge relationship better, a view of code 4.3, shows that this logic follows seeing that nodes that are related in the edge tables share the same attribute value in common in the xml snippet.

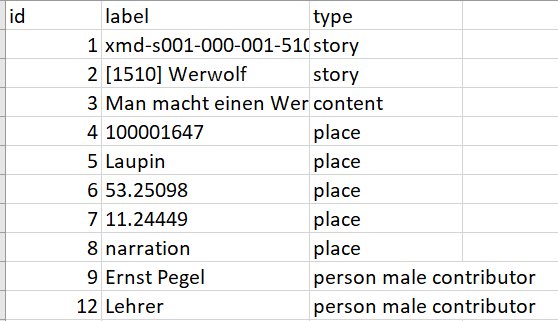


Table 41. wossidia node file

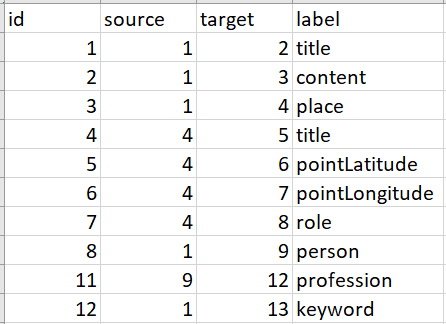
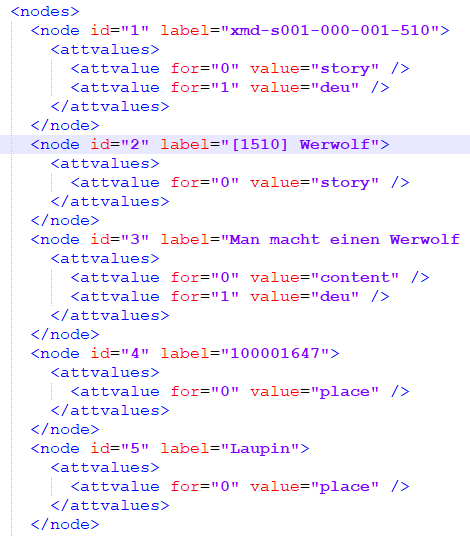
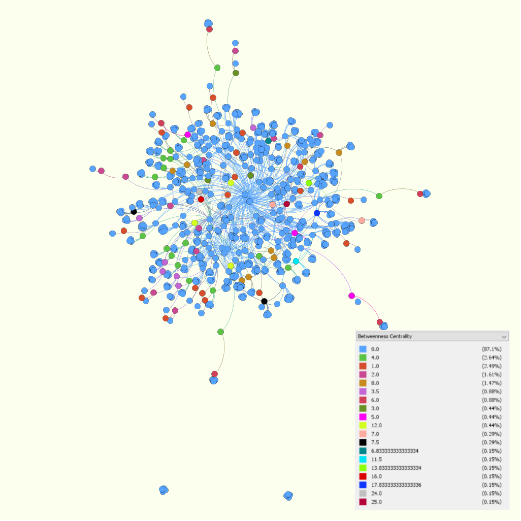


Table 41. wossidia edge file



Code 4.3

Since the Wossidia dataset is a graph data, in order to get an in-dept analysis of it, it is important to visualize it using a data visualization tool like Gephi. Figure 4.4 represents a network obtained when the dataset is imported into the Gephi tool. The Partitioning of this network has been done based on the betweenness centrality measure and color code has also been applied. Statistics obtained based on this shows that majority of the nodes has an interconnection as can be seen in the attached table beside the network, 87.1% of the nodes are somehow linked together, this is also evident in the visualization of the network. Other statistics obtained from the network has also been tabulated in table 4.4



[1] A. C. Schering, I. Bruder, C. Schmitt, H. Meyer, and A. Heuer. 2007. Towards a digital archive for handwritten paper slips with ethnological contents. In Proceedings of the 10th international conference on Asian digital libraries: looking back 10 years and forging new frontiers (ICADL). Springer-Verlag, Berlin, Heidelberg, 61–64.