General Approach of Buster

The general approach of BUSTER (Bremen University Semantic Translator for Enhanced Retrieval) follows the hybrid ontology approach described in section 2. This means that the overall architecture is based on annotated information sources, which are linked or can be found with an ontology-based retrieval mechanism. BUSTER can be looked at as a middleware, which can be used by various applications such as e-commerce programs, GIS-applications etc. It provides two subsystems, one acts as intelligent search engine, the other can be used for information integration or semantic translation. This general approach has been also discussed in [127, 133, 129, 130, 141, 132, 110].

3.1 Requirements

We consider the needs and future research lines of the Semantic Web community and can define the following overall requirements of the system:

Firstly, we would like to set up the system with an intelligent search component. This means, that the system should be able to find information sources with intelligent search methods described in section 1.3. Integration and/or translation of the found information is another important task, which includes the integration of information on the concept level and also an optional context transformation on the data level. This is just one of the essential needs for the Semantic Web to become successful. However, this implies that the data on the Web have to be annotated with background knowledge. Moreover, this background knowledge has to have some kind of formality to provide full support for both the retrieval and the integration process.

Secondly, the user should be able to query the Web with more than just the terms they are seeking. An important feature is the search for spatial terms or concepts. An example is someone looking for accommodation in a certain place. This place should also include colloquial terms rather than (x,y)-coordinates, which are used within monolithic GIS. Further, the inclusion of GIS into our approach would involve the download of a huge amount of data.
given the fact that GIS “inferences” run on polygons. These polygons however, are usually of high resolution and therefore contain a lot of data. GIS are normally used for planning purposes and are run by official departments. We argue that many Semantic Web applications do not need data of this high resolution. In addition, using the GIS data on a more qualitative level would be of high value. One reason for this is the amount of data traffic (in terms in bytes) on the Web.

Another important aspect is the possibility to look for temporal terms on the Web. An extension of the former example clarifies this: we should be able to look for accommodation at certain places during a certain time. These temporal terms should also be colloquial such as “Easter vacation 2003”. Currently, the W3C offers time specification with two exact time stamps. We will see later (section 6) that this does not fulfill the needs for the Semantic Web.

### 3.2 Conceptual Architecture

The BUSTER architecture provides an integrated solution for the problem of information retrieval and integration. We take into account all three levels of integration (cf section 1.4) combining several technologies including: standard markup languages, mediator systems, ontologies and knowledge-based classifiers. This holds particularly true for the terminological part of the BUSTER system, however, the overall approach of using ontologies and a common vocabulary also applies for the spatial and temporal part.

Figure 3.1 gives an overview about the BUSTER architecture. In general, the architecture can be divided in two distinct phases: an acquisition phase and a query phase.

During the acquisition phase, all desired information for providing a network of integrated information sources is acquired. This includes the acquisition of a Comprehensive Source Description (CSD) (see section 3.3 below) of each source together with the Integration Knowledge (IK), which describes how the information can be transformed from one source to another.

In the query phase, a user or an application (e.g., an e-commerce application, a GIS or a user searching for information on the Web) formulates a query which implies to an integrated view of sources. Several specialized components in the query phase use the acquired information, i.e., the CSD’s and IK’s, to select the desired data from several information sources and transform it into the structure and the context of the query.

All software components in both phases are associated to three levels: the syntactic, the structural and the semantic level. The components on each level deal with the corresponding heterogeneity problems. The components in the query phase are responsible for solving the corresponding heterogeneity problems whereas the components in the acquisition phase use the CSD’s from