Enterprises use a wide range of heterogeneous information systems in their business activities such as Enterprise Resource Planning (ERP), Customer Relationships Management (CRM) and Supply Chain Management (SCM) systems. In addition to the large amounts of heterogeneous data produced by those systems, external data is an important resource that can be leveraged to enable taking quick and rational business decisions.
Classic Business Intelligence (BI) and even the newer Agile Visualization tools focus much of their selling features on attractive and unique visualizations. Preparing data for those visualizations still remains the far most challenging task in most BI projects large and small. Self-service data provisioning aims at tackling this problem by providing intuitive datasets discovery, data acquisition and integration techniques to the end user.

The goal of this thesis is to provide a framework that enables self-service data provisioning in the enterprise. This framework empowers business users to search, inspect and reuse data through semantically enriched datasets profiles.

Publicly available datasets contain knowledge from various domains such as encyclopedic, government, geographic, entertainment and so on. The increasing diversity of these datasets makes it difficult to annotate them with a fixed number of pre-defined tags. Moreover, manually entered tags are subjective and may not capture their essence and breadth. We propose a mechanism to automatically attach meta information to data objects by leveraging knowledge bases like DBpedia and Freebase which facilitates data search and acquisition for business users.

In many knowledge bases, data entities are described with numerous properties. However, not all properties have the same importance. Some properties are considered as keys for performing instance matching tasks while other properties are generally chosen for quickly providing a summary of the key facts attached to an entity.
Business users may want to enrich their reports with these data entities. To facilitate this, we propose a mechanism to select what properties should be used when augmenting extra columns into an existing dataset or annotating instances with semantic tags.

Linked Open Data (LOD) has emerged as one of the largest collections of interlinked datasets on the web. In order to benefit from this mine of data, one needs to access to descriptive information about each dataset (or metadata). This metadata enables dataset discovery, understanding, integration and maintenance. Data portals, which are datasets' access points, offer metadata represented in different and heterogeneous models. We first propose a harmonized dataset model based on a systematic literature survey that enables complete metadata coverage to enable data discovery, exploration and reuse by business users. Second, rich metadata information is currently very limited to a few data portals where they are usually provided manually, thus being often incomplete and inconsistent in terms of quality. We propose a scalable automatic approach for extracting, validating, correcting and generating descriptive linked dataset profiles. This approach applies several techniques in order to check the validity of the metadata provided and to generate descriptive and statistical information for a particular dataset or for an entire data portal.

Traditional data quality is a thoroughly researched field with several benchmarks and frameworks to grasp its dimensions. Ensuring data quality in Linked Open Data is much

more complex. It consists of structured information supported by models, ontologies and vocabularies and contains queryable endpoints and links. We propose an objective assessment framework for Linked Data quality based on quality metrics that can be automatically measured. We further present an extensible quality measurement tool implementing this framework that helps on one hand data owners to rate the quality of their datasets and get some hints on possible improvements, and on the other hand data consumers to choose their data sources from a ranked set.

Finally, the Internet has created a paradigm shift in how we consume and disseminate information. Data nowadays is spread over heterogeneous silos of archived and live data. People willingly share data on social media by posting news, views, presentations, pictures and videos. We propose a service that brings relevant, live and archived information to the business user. The key advantage is an instantaneous access to complementary information without the need to search for it. Information appears when it is relevant enabling the user to focus on what is really important.