**ABSTRACT**: The increasing number of functionally similar services requires the existence of a non-functional properties selection process based on the Quality of Service (QoS). Thus, in this article authors focus on the provision of a QoS model, an architecture and an implementation which enhance the selection process by the annotation of Service Level Agreement (SLA) templates with semantic QoS metrics. This QoS model is composed by a specification for annotating SLA templates files, a QoS conceptual model formed as a QoS ontology and selection algorithm. This approach, which is background compatible, provides interoperability among customer-providers and lightweight alternative. Finally, its applicability and benefits are shown by using examples of Infrastructure services.