SWAM project assignment 2020-2021

I18N-Store: a RESTful application featuring full-text search on localized items

Student: Francesco Areoluci

Credits: 9 CFUs

Abstract

Hibernate Search is an enabling technology for full-text search over ORM entities for a Java Application. This library allows the developer to add search functionalities with an easy-to-use interface. Full-text search is implemented using Lucene, for local indexing, and Eleasticsearch for remote indexing. This project implement Hibernate Search functionalities in an application which represent an online shop. The shop features the localization of its products: a customer will receive products localized in his/her locale.

1 Introduction

Hibernate Search is an enabling technology for indexing and full-text search on Hibernate applications. In particular, it allows to extract data from hibernate ORM entities and push it to local or remote indexes. Hibernate Search is particularly useful for applications where SQL-based searches are not suited, such as full-text and geolocation searches. The main difference with traditional search is that the stored text is not considered as a single block of text, but as a collection of tokens (words).
In order to do that, Hibernate Search consists of an indexing component, which associates indexes to entities, and an index search component, which allows to search through indexes. These services are offered by Apache Lucene, an high-performance, full-featured text search library written in Java. Indexes are created and updated each time an entity is inserted, updated or removed from the database. Once the index is created, entity search can be accomplished without dealing with the underlying Lucene infrastructure.
Moreover, remote indexing and searching can be accomplished using Elasticsearch, a distributed search engine. These search engines are based on the concept of inverted indexes: a dictionary where the key is a token found in a document and the value is the list of identifiers of every document containing the token. A search involves the following steps:

- Extract tokens from the input query;
- Lookup tokens in the index to find matching documents;
- Aggregrate results of the lookup to produce a list of matching documents

The project's application has been developed by integrating the Hibernate Search features. In particular, the application represents the backend of an online shop which let the users to purchase products and search for them through query on their name and/or

description. Products fields, such as name and description, are localized in multiple locales: this project features the italian and english localization. Products are displayed to customer in his/her configured locale: an italian customer will get products which fields are localized in italian.
Through this application, Hibernate Search functionalities can be used on product's localized attributes: each product has fields available in multiple languages and those localized versions belong to the same product. These fields in multiple languages will be used by Hibernate Search to retrieve matching entities.

1.1 About Internationalization and Localization

Localization, often referred as *l10n* where 10 is the number of letters between *l* and *n*, is the process of adapting a product, an application or a textual document to a specific country or region. Localization is not limited to textual translation but it includes a variety of adaptations, such as:

- adapting graphics;
- adapting to local currencies;
- adapting date/time format;
- adapting addresses format

The process of localization in a product or application can be easily enabled through a process of internationalization.
Internationalization, often referred as *i18n* where 18 is the number of letters between *i* and *n*, is the process of design and development of a product or application that **enables easy localization**. As an example, an application architectural design that enables an easier localization of the content of the application, is a process of internationalization.

2 Hibernate Search Architecture

This chapter offers a brief view of the Hibernate Search architectures and modules. In particular, Hibernate Search is composed by the following modules and functionalities:

- **Backend**: The backend module abstracts the full-text search engine, by implementing indexing and searching interfaces. Hibernate Search provides two abstraction: Lucene backend and Elastisearch backend. The backend can be configured using Hibernate configuration.
- **Mapper**: This module allows to map the user model to an index model. In particular, Hibernate Search mapper offers the indexing of Hibernate ORM entities. This can be done through annotations or programmatic API.
- **Mass indexer**: Thanks to the mass indexing, Hibernate Search can rebuild indexes starting from pre-exisiting data stored on database.
- **Automatic indexer**: This allows to keeps indexes in sync with a database. Each time an entity is added, updated or removed, the mapper detects the changed and the index is updated.
- **Searching**: This is how Hibernate Search provides ways to query the index. The search involves the creation of a query, the token extraction and filtering, the search in the index and the entities retrieval.

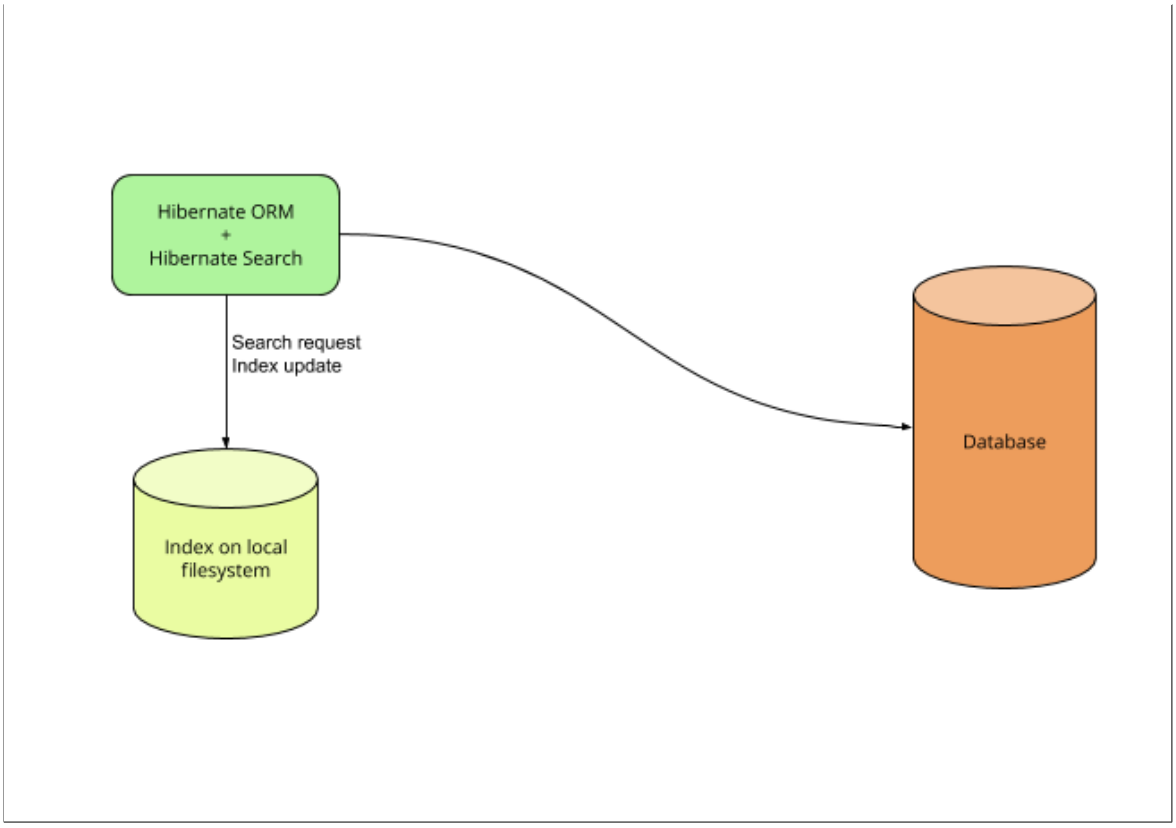

Figure 1: Hibernate Search Lucene Backend

3 Application requirements

The design of the application has started through the definition of its requirements. To do that, requirements have been divided in four categories:

- **FR**: Functional Requirements
- **NFR**: Non-Functional Requirements
- **DR**: Domain requirements
- **C**: Project Constraints

3.1 Functional Requirements

- **FR1**: The system must allow to manage the available products
- **FR2**: Each product must contains the following fields:
  - **FR2.1**: Product Name
  - **FR2.2**: Product Description
  - **FR2.3**: Price
  - **FR2.4**: Manufacturer
  - **FR2.5**: Product category
- **FR3**: The system must have the localization feature: products must be localized in multiple langauges (it, en). Localization feature must be used on the following fields:
  - **FR3.1**: Product Name
  - **FR3.2**: Product Description
  - **FR3.3**: Price
  - **FR3.4**: Product Category
- **FR4**: The system must allow the visualization and purchase of the products.
  - **FR4.1**: The product visualization must be localized: a product must be returned to the user with fields specified in FR3 properly localized.
- **FR5**: The system must features the search on products.
- **FR6**: The system must features the management of two types of user account
  - **FR6.1**: Administrator account - Responsabilities: Product management and research, as specified in FR1 and FR5
  - **FR6.2**: Customer account - Responsabilities: Product research and purchase, as specified in FR4 and FR5

3.2 Non-Functional requirements

- **NFR1** (Security Requirement): Application access must be managed through user authentication for all the account types described in FR6.
- **NFR2** (Implementation Requirement): The purchase functionality, as specified in FR4, must be prototyped as following:
  - **NFR2.1**: The user must have a shopping cart. The user can add and remove products to/from the cart;
  - **NFR2.2**: Once the user has added to the cart the desired prodcuts, he/she can proceed to the checkout. This operation will move the cart products to a shopping list. The cart will be cleared.

3.3 Domain Requirements

- **DR1**: The product management, as specified in FR1, must allow to add, remove and edit the products (CRUD operations)
- **DR2**: The products search, as specified in FR5, must be based on query keywords and must be used to search on products' name and description.
- **DR3**: A customer account must have an associated locale in order to implement the product visualization feature.

3.4 Project Constraints

- **C1**: The application must be developed using Java EE, with JPA and CDI technologies.
- **C2**: The Persistence layer must be managed using Hibernate.
- **C3**: The persistence must be managed by DBMS MariaDB.
- **C4**: The search layer must be manager through Hibernate Search library.
- **C5**: Application services must be exposed via REST API using JAX-RS technology.
- **C6**: Package management must be managed through Maven.

4 Application Design

4.1 Use Cases design

Using the previously described requirements, the use cases for the two types of account have been designed. The following image represents the use case diagram that has been created to show which operation each type of user can perform using the application.

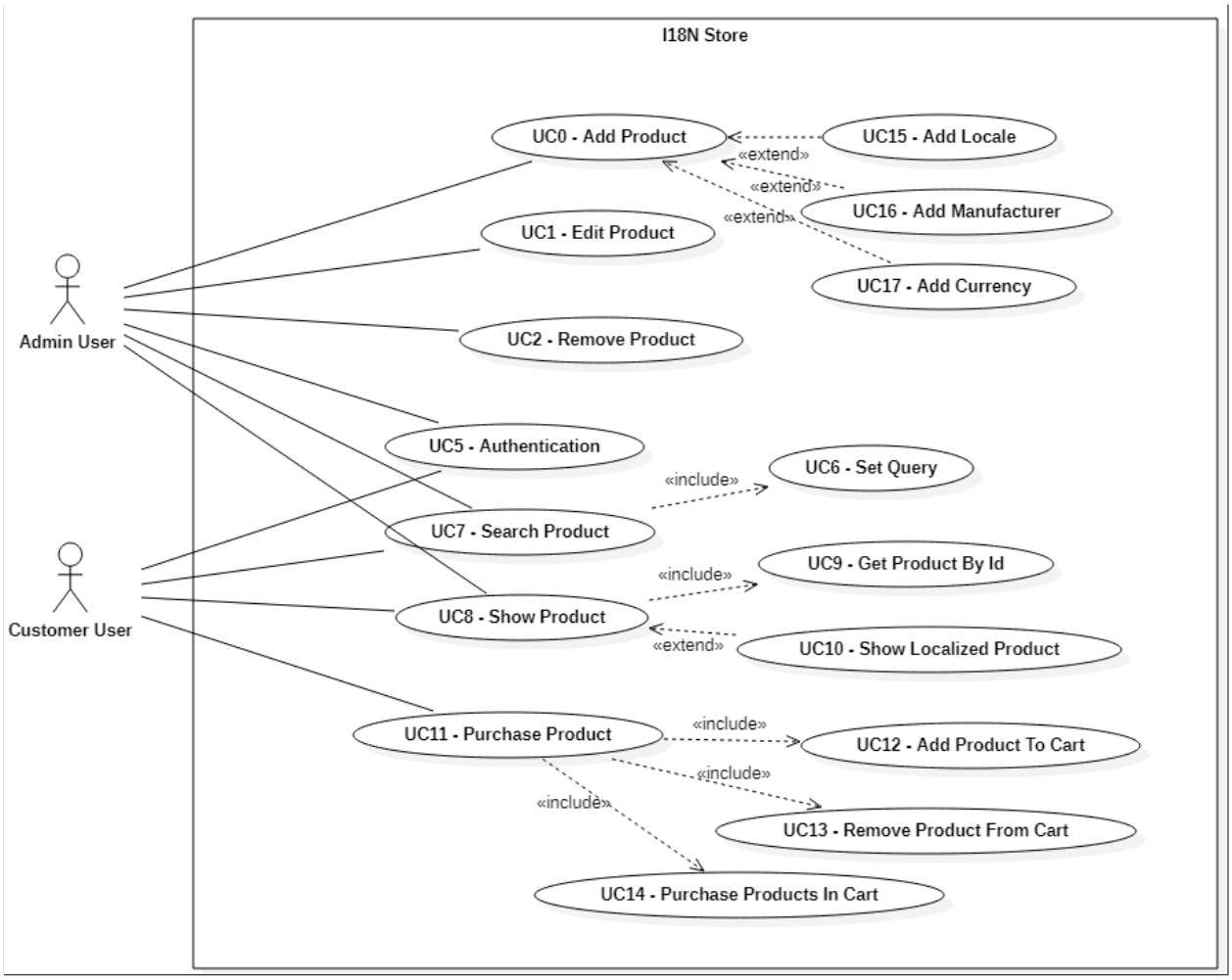

Figure 2: Use Case Diagram

As we can see from the diagram:

- an administrator user can interact with the system through insert, update and delete of products. To accomplish that, he/she can add manufacturer, locales and values;
- a customer user can interact with the system through (localized) visualization and purchase of products;
- common use cases between the two users are authentication, search and visualization of products.

4.2 Class design

Once the use cases have been designed, the application has been modeled through the UML class diagram. The application has been divided in packages, one for each logic domain. The diagram exposes the following packages:

- **Domain Model**: contains the entities that represents the application domain;
- **Translation Model**: contains the entities responsible of the field translation handling;
- **Controllers**: contains the available controllers (endpoints);
- **Data Access Objects**: contains the entities responsible for DBMS interaction and instantiation of domain model entities;
- **Data Transfer Objects**: contains the entities used to transfer data from endpoints to client and vicevers, in JSON format;
- **Security**: contains the entities responsible of user authentication and authorization management.

The following image shows the class diagram of the entire application.

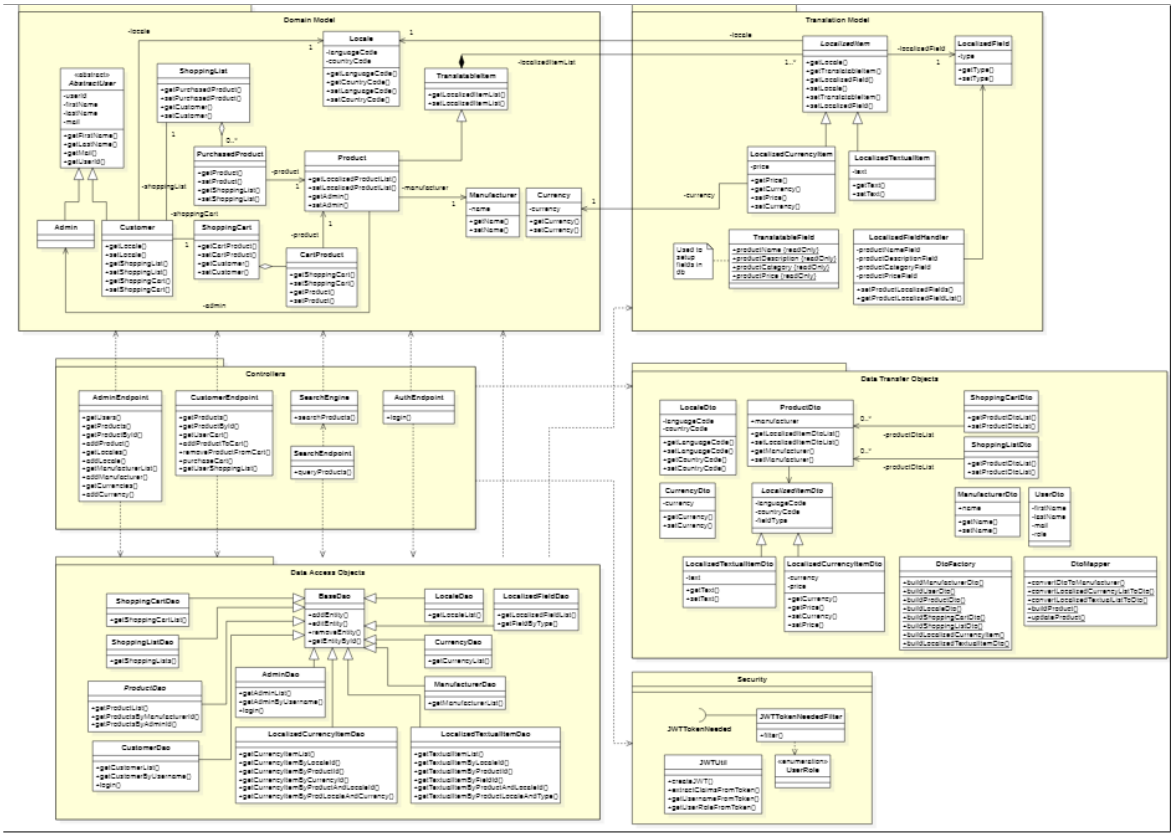

Figure 3: Complete class diagram

Each package will be examinated in the following sections.

4.2.1 Domain Model

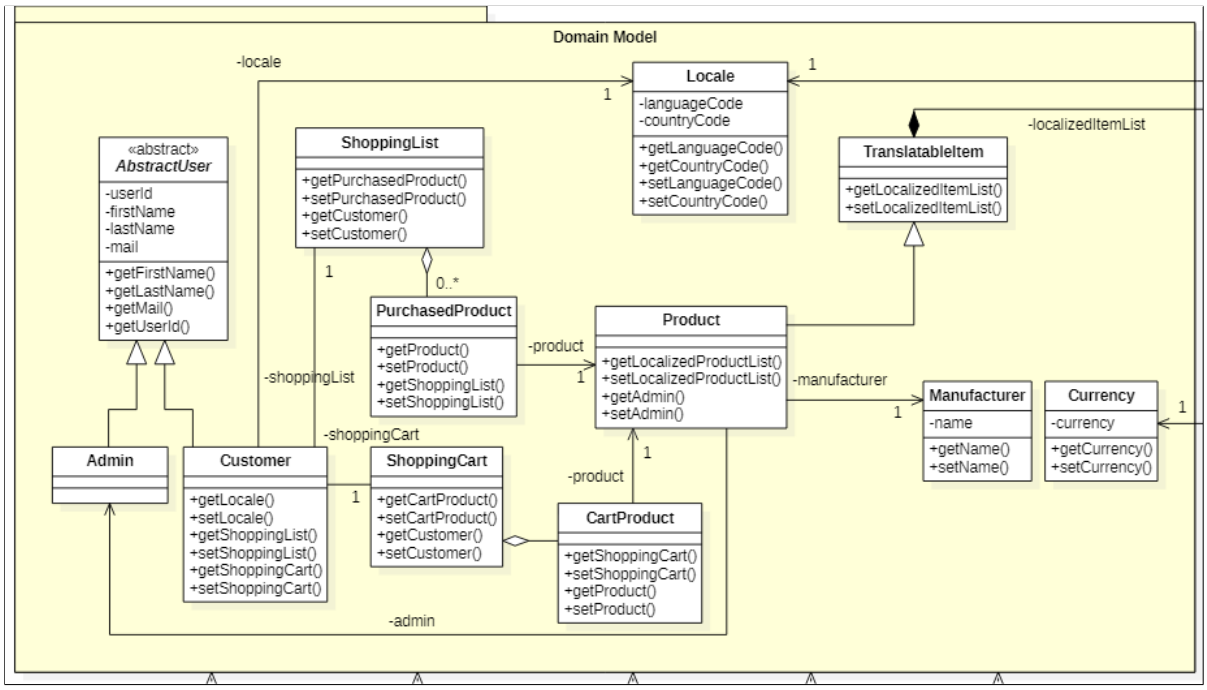

Figure 4: Domain Model Class Diagram

The domain model represents the base entities of the shop. It contains mainly POJO entities. In particular, it features the following entities:

- the two types of user: **Admin** and **Customer**;
- the **Product** entity, which instance is inserted by an Admin, has a **Manufacturer** and that subclasses a **TranslatableItem**. This entity has a relationship with the LocalizedItem entity of the Translation Model: each entity that need a translation of some fields can subclass the TranslatableItem entity;
- the **ShoppingCart** and the **ShoppingList**. These entities are associated respectively to the product in cart (**CartProduct**) and to the **PurchasedProducts**;
- the **Locale** entity, which is associated to Customer and LocalizedItems;
- the **Currency** entity, which represent te application available currencies.

These entities are instantiated and persisted by the controllers and the Data Access Objects. The localization of the fields of the Product entities is managed through the Translation Model.

4.2.2 Translation Model

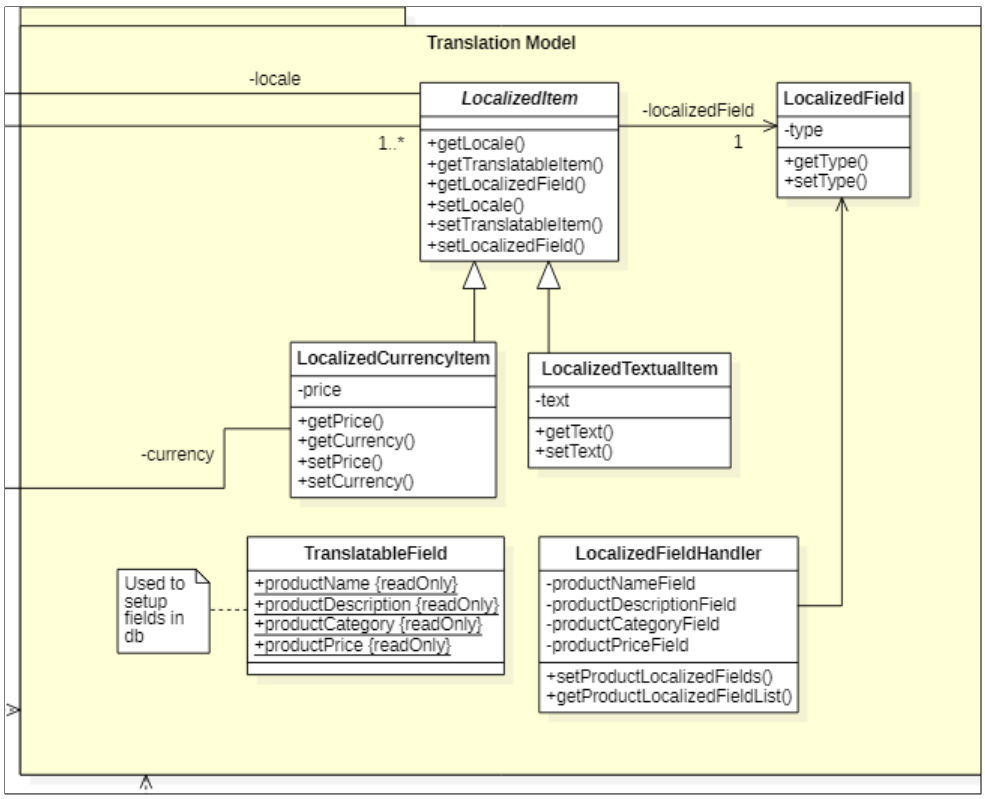

Figure 5: Translation Model Class Diagram

This package is used to manage the entity translations. Each entity that wants the localization feature for certain fields has to mantain an association with one or more **LocalizedItem**: this superclass is associated with a certain **Locale** and a certain **LocalizedField**.
The LocalizedField indicates for which fields the localization takes place (e.g product name, product description, etc.). Thus, a LocalizedItem is referred to a TranslatableItem, a Locale and a certain LocalizedField.
Two types of subclasses for the LocalizedItem have been created: **LocalizedTextualField** and **LocalizedCurrencyField**, which, respectively, can be used to represent a textual field or a price field (along with a currency).
The **TranslatableField** entity exposes static strings that can be used to persists the allowed localization fields.
The **LocalizationFieldHandler** entity can be used to manage the allowed localization fields: it will populate its fields with the persisted localized fields and expose them for entities that wants to translate a certain item.

4.2.3 Controllers

The controllers can manipulate the domain model entities through services that exposes functionalites to the application's clients. In particular, the controllers exposed by this layer offers:

- authentication interface: **AuthEndpoint**;

- administration interface: **AdminEndpoint**;
- customer interface: **CustomerEndpoint**;
- search interface: **SearchEndpoint** which uses the entity responsible for the product search (**Search Engine**)

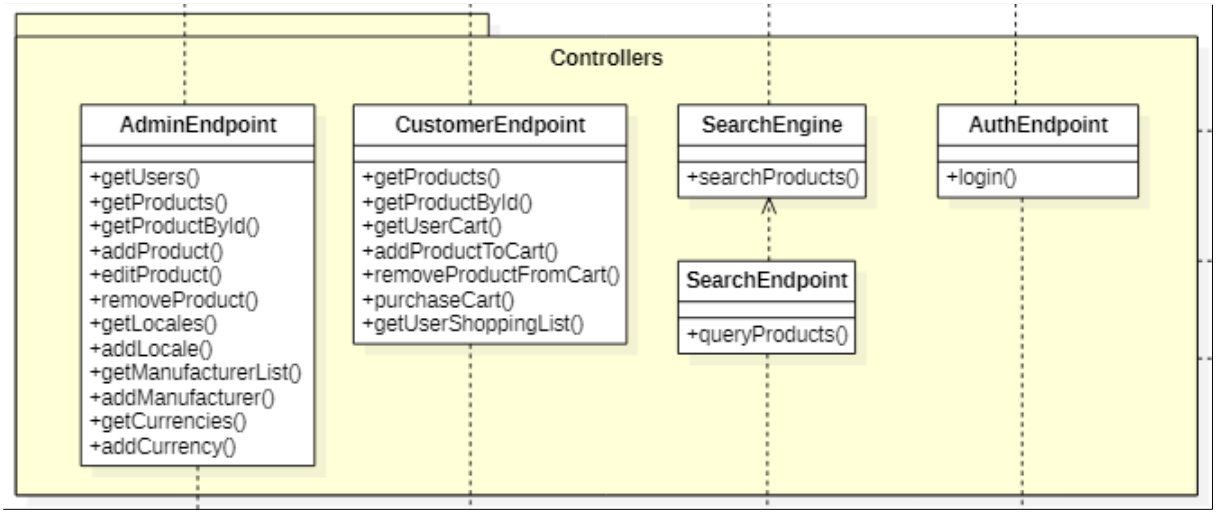

Figure 6: Controller Class Diagrams

The controllers interact with the domain model entities using the Data Access Objects (DAO) exposed by the dedicated layer. These controllers are exposed to the clients as REST interfaces, by usign the JAX-RS library. The information exchange between this layer and the clients is managed through JSON objects, which are created from base entities exposed by the Data Transfer Object layer.

Authentication and authorization are managed using **Json Web Token** (JWT). This type of token is self-contained and allow a stateless user authentication thanks to the signature emitted by a certificate authority. In particular, the authentication management in this layer is accomplished thanks to the annotation exposed in the Security layer.
JWT will be described more in detail in section **4.2.6**.

4.2.4 Data Access Objects

Data Access Object entites implement DBMS access and persistence functionalities, and offer them to the Controllers layer. The common functionalities between all DAO entities are contained in **BaseDao** and are the followings:

- Search entities by identifier
- Persist a new entity
- Update a persisted entity
- Delete a persisted entity

The other entities implent specific operations based on the referenced Domain Model entity.

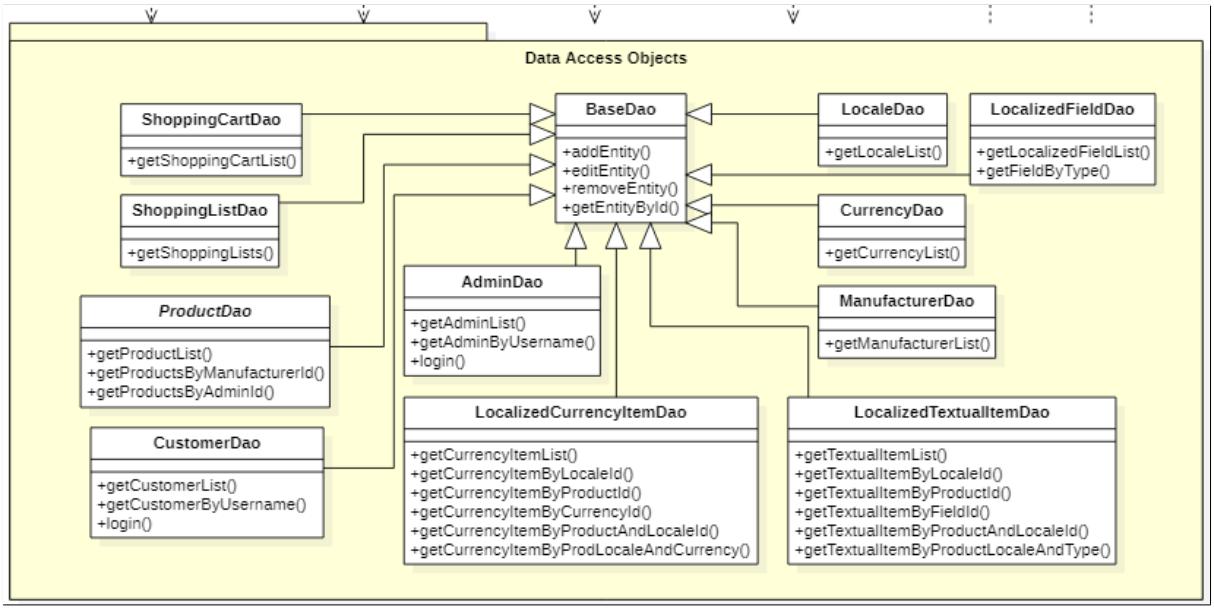

Figure 7: DAO Class Diagrams

4.2.5 Data Transfer Objects

Data Transfer Objects are POJO entities that allow to build object that can be sent or received on REST calls as JSON. They are used by some endpoints of the Controllers layer to manage input and response data. This layer has been used in order to abstract the modeling of messages between the Controllers layers and the clients from the Domain Model entities.

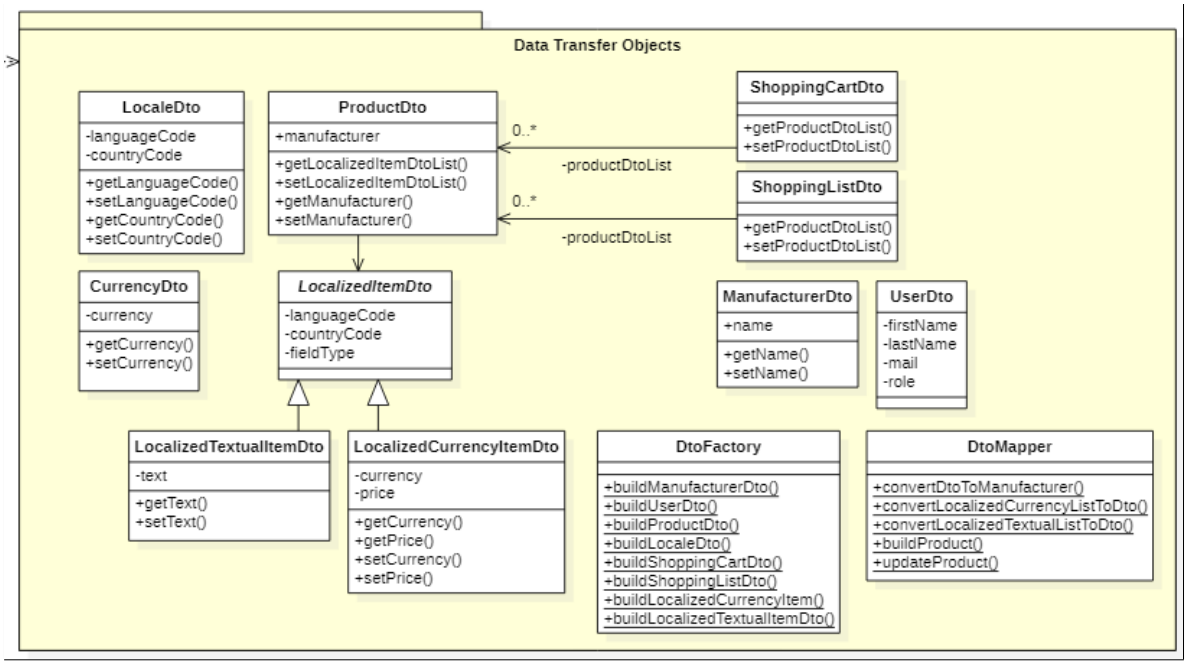

Figure 8: DTO Class Diagrams

4.2.6 Security

This layer implements authentication and authorization features by using Json Web Tokens (JWT). As said, JWT are self-contained tokens that enables stateless authentication thanks to the certificate signature: a tampering of the token will invalidate the signature.
A Json Web Token is composed of three parts:

- **Header**: Contains the signing algorithm;
- **Payload**: A set of keys/values which store application data. Keys are called *claims*;
- **Signature**: Created by passing header and payload to the signing algorithm.

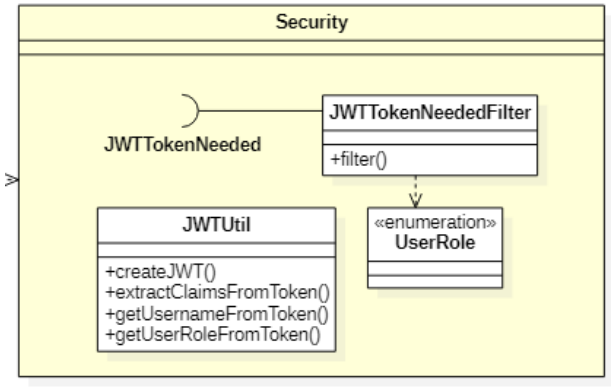

Figure 9: Security Class Diagrams

Each part of a JWT is encoded as Base64 and the three part are concatenated using the "." character.
Tokens are signed using HMAC + SHA256. For simplicity reasons, tokens emission and validation are handled by the application instead of using a separated authentication server.
These tokens have the following payload structure:

```
{
  "subject": "mario.rossi@example.com",
  "issuedAt": 1624223874036,
  "userRole": "ADMIN",
  "exp": 1624224474,
  "issuer": "i18n-store"
}
```

The user role has beeen included in the token claims in order to handle the authorization on exposed enpoints. This way, administration endpoint are available only for users with ADMIN role, while customer endpoints are available only for user with CUSTOMER role.

Note: Tokens are not encrypted. They must be exchanged over an SSL/TLS layer to guarantee token encryption, for example by using HTTPS.

4.3 REST Endpoints

The following URI are the application's exposed endpoints.

- Authentication endpoint (**/api/auth/**):
  - **/api/auth/login**: User authentication, sent JWT to user if given username and password are correct
- Admin endpoints (**/api/admin/**):
  - **/api/admin/users**: Show all the users
  - **/api/admin/products**: Show all the products, along with all localizations
  - **/api/admin/products/{id}**: Show specified product, along with all localizations
  - **/api/admin/products/add**: Add a product
  - **/api/admin/products/edit**: Edit a product
  - **/api/admin/products/remove/{id}**: Remove a product
  - **/api/admin/locales**: Show all configured locales
  - **/api/admin/locales/add**: Add a locale
  - **/api/admin/manufacturers**: Show all manufacturer
  - **/api/admin/manufacturers/add**: Add a manufacturer
  - **/api/admin/currencies**: Show all currencies
  - **/api/admin/currencies/add**: Add a currency
- Customer endpoints (**/api/customer/**):
  - **/api/customer/products**: Show all products, localized in user locale
  - **/api/customer/products/{id}**: Show specified product, localized in user locale
  - **/api/customer/shopping-cart**: Show user shopping cart
  - **/api/customer/shopping-cart/add/{id}**: Add specified product to user shopping cart
  - **/api/customer/shopping-cart/remove/{id}**: Removed specified product from the user shopping cart
  - **/api/customer/shopping-cart/checkout**: Purchase all the product in shopping cart and move them to the shopping list
  - **/api/customer/shopping-list**: Show user shopping list
- Search endpoint (**/api/search/**)
  - **/api/search/products/{query}**: Search products according to their name and description with a query (keyword separated by "+" character)
  - **/api/search/products/similar-to/{id}**: Search products similar to the one specified by the given identifier

4.4 Sequence Diagrams

Through sequence diagram, specific actions of an use case can be modeled in order to describe what messages are exchanged between the entities to accomplish the operation. The following sequence diagrams have been created:

- Add Product
- Show Product
- Add Product To Cart
- Purchase Products

These diagrams show the execution procedures for some operations of the specified use cases in section **4.1** and that can be implemented with the modeling described in section **4.2**.

4.4.1 Add Product

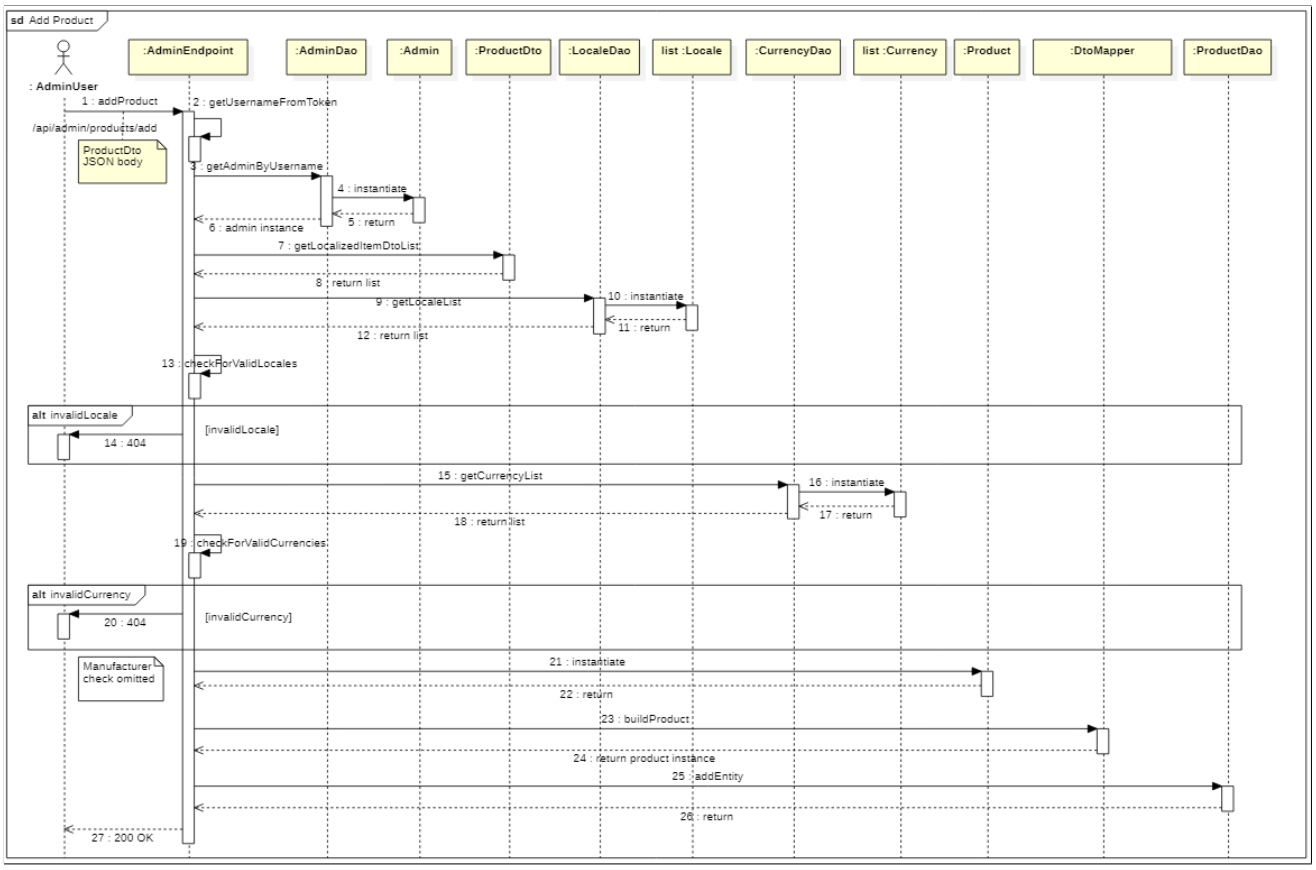

Figure 10: Add Product Sequence Diagram

4.4.2 Show Product

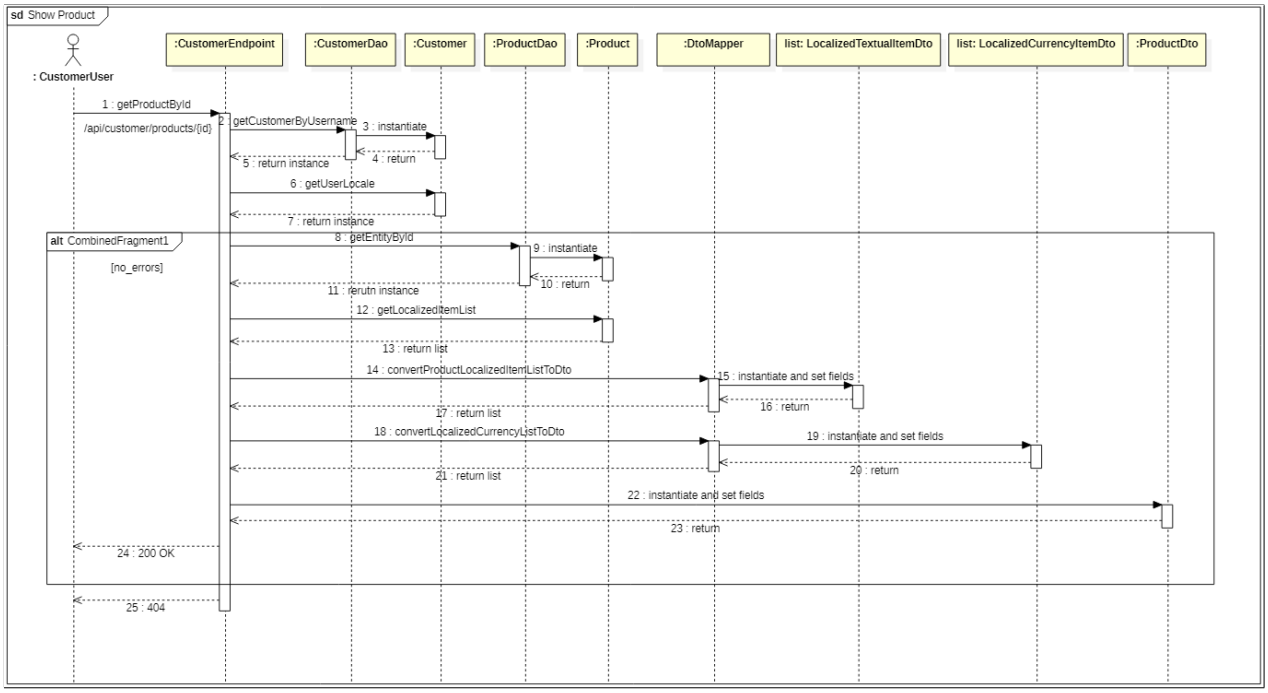

Figure 11: Show Product Sequence Diagram

4.4.3 Add Product To Cart

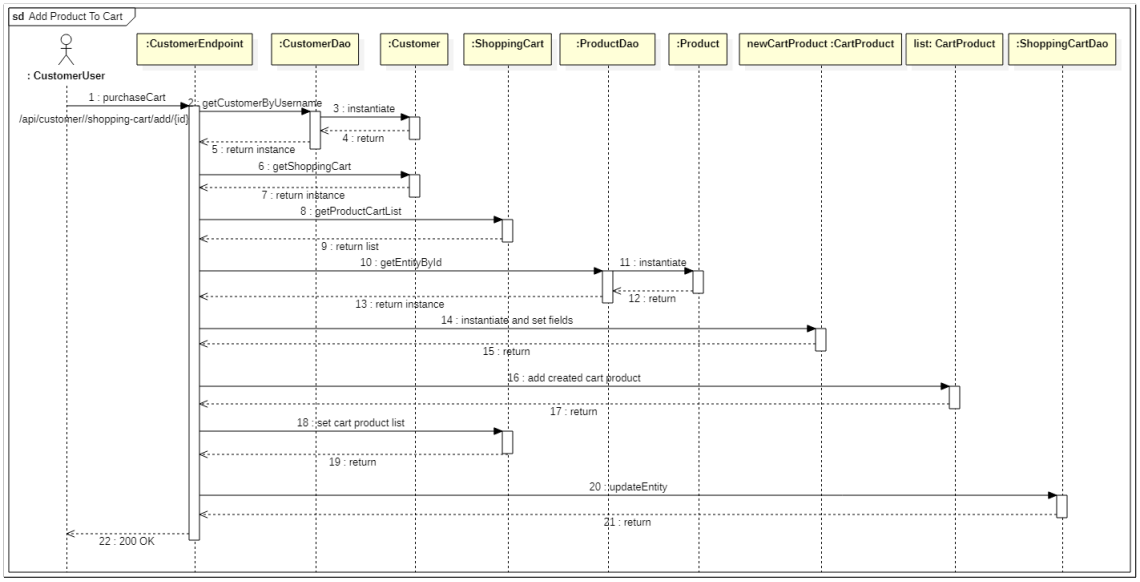

Figure 12: Add Product To Cart Sequence Diagram

4.4.4 Purchase Products

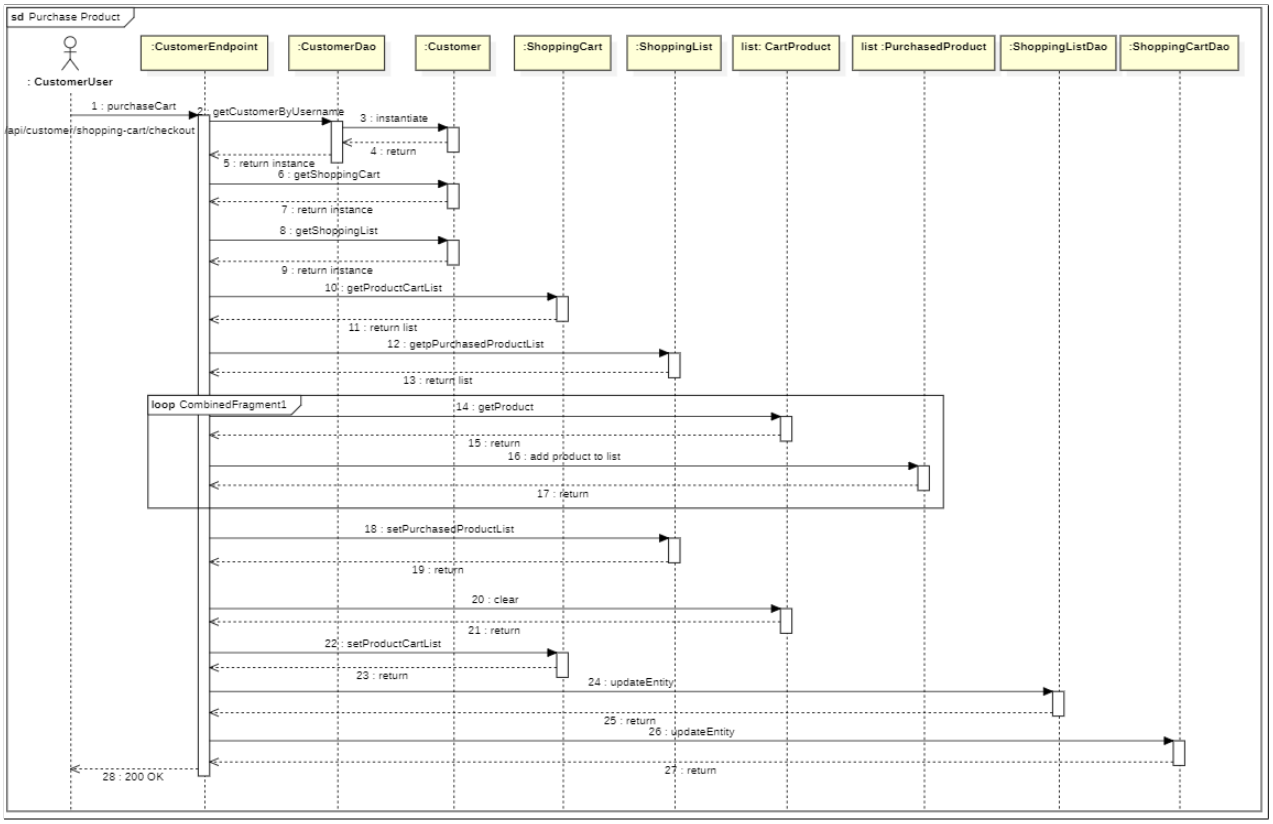

Figure 13: Purchase Products Sequence Diagram

4.5 ER Model

The following image shows the implementation of the relation database model on which the application is based on:

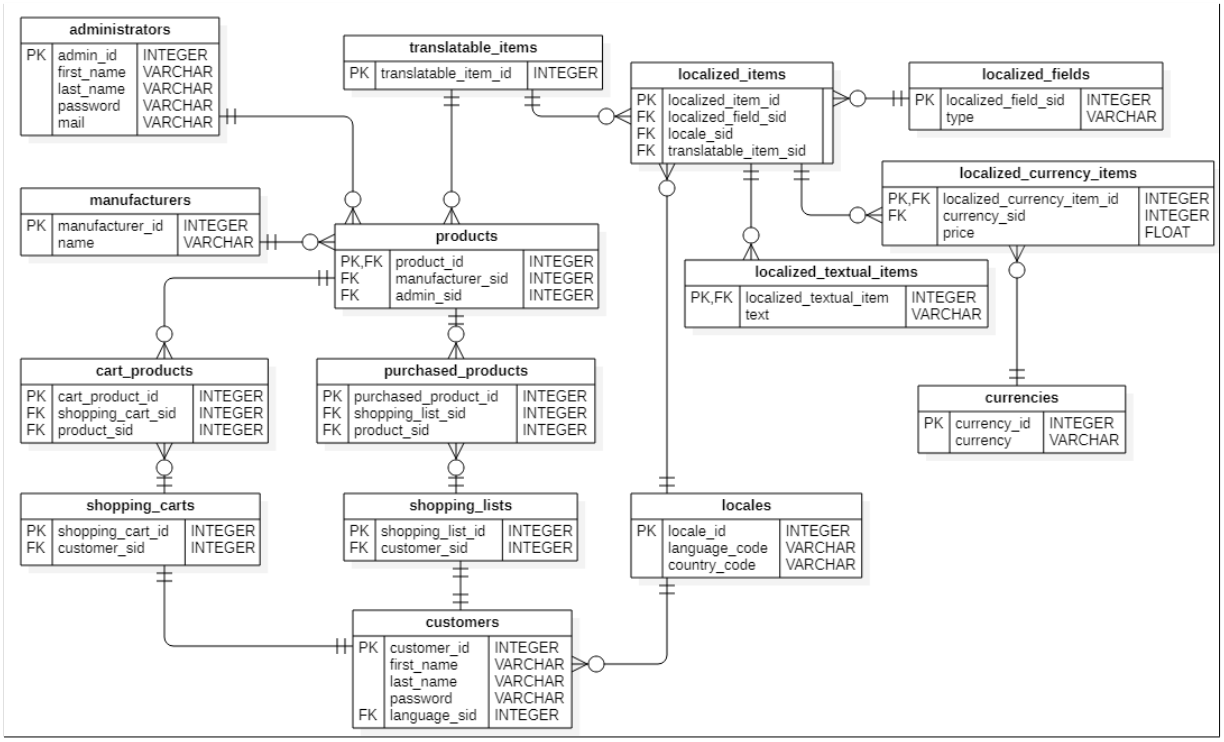

Figure 14: DB ER Model

The model is based on the following entities:

- **administrators**: admin users that manage the products as described in use cases **UC0**, **UC1** and **UC2**;
- **customers**: customer users, they manages a **shopping cart** and a **list of purchased products**. They are associated to a specific **locale**;
- **cart products**: associated to a specific **shopping cart** and to a specific **product**;
- **purchased products**: associated to a **shopping list** and to a **specific product**;
- **products**: available products, associated to a **manufacturer** and inserted by an **administrator**. This table associated to the **translatable items** table (Product's superclass);
- **localized items**: localizations are separated from the **translatable items** table and each localization is associated to a **locale** and a **localized field**. This way, if a locale is added to support the application, it will result in a new row in this table for each product, instead of new fields added in products table. Moreover, it is associated with **textual items** and **currency items** tables (subclasses). A currency item is associated to a certain **currency**, to express the product price.

5 Packages Testing

In order to verify and test the developed code, a test suite has been created. This suite allows to test the packages in an isolated way, using JUnit, and to execute integration test of all packages, using Postman.
Through JUnit, test cases have been created for the following packages:

- **Model**
- **Dao**
- **Dto**
- **Security**

Controllers tests and integration tests have been created through Postman. Thanks to Postman, a collection of REST calls has been created, in order to test all the exposed endpoints. Thus, correctness of returned values and error handling can be easily verified by running the collection.

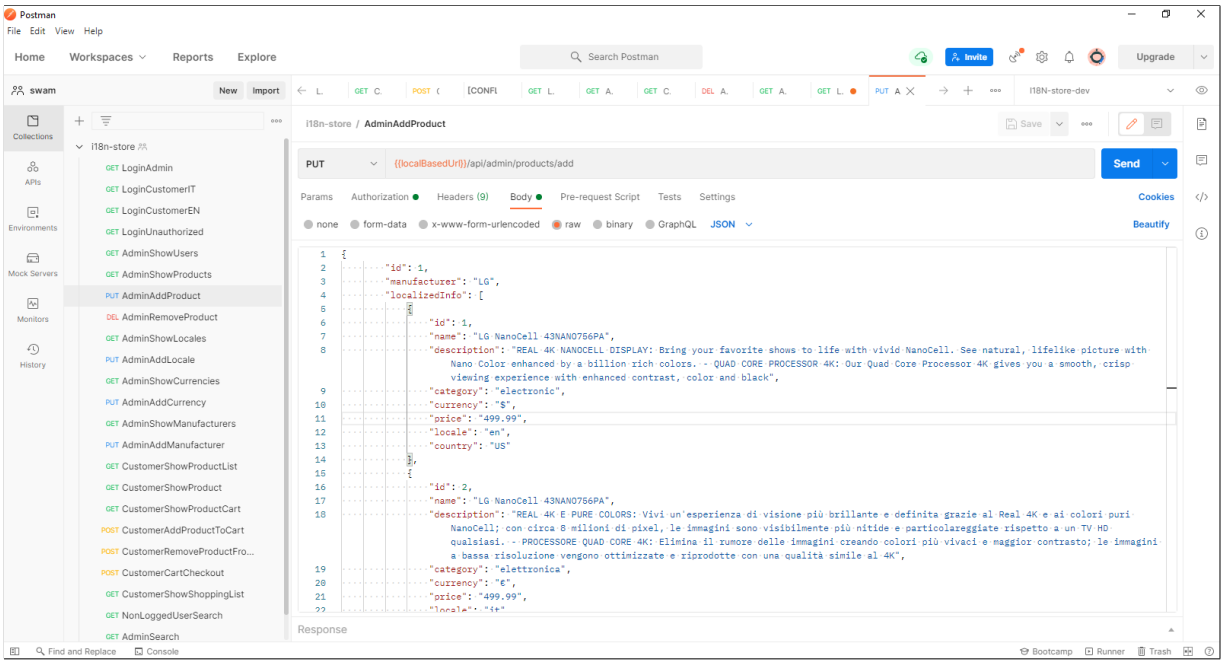

Figure 15: Postman

6 Full-Text Search

Product search and retrieval have been accomplished through integration of Hibernate Search functionalities. Using the version 5.10.7 of this library, the following functionalities have been implemented:

- Startup indexing
- Indexing annotations
- Standard Search
- Standard AND Search
- Fuzzy Search
- Phrase Search
- Similarity Search

These implementations will be briefly described in the following sections.

6.1 Startup Indexing

Hibernate Search index is used to perform fast searching of entities by storing keywords that belongs to certain documents. The index can be mantained in filesystem or can be temporary stored in memory. In both cases, at startup, the Hibernate Search indexer component must be initialized. This can be done as:

```
FullTextEntityManager fullTextEntityManager = Search
                                .getFullTextEntityManager(entityManager);
fullTextEntityManager.createIndexer().startAndWait();
```

This way, persisted entities that are not already indexed will be stored in the index and thanks to the automatic indexer the indexes will be kept in sync with the database objects.

6.2 Indexing annotations

Entities that should be considered for index building and retrieval must be annotated with Hibernate Search annotations.
An entity that should be indexed must be annotated with the keyword **@Indexed** and its searchable fields must be annotated with **@Field**. Moreover, entities can be defined to be embedded into an indexed entity with the keyword **@IndexedEmbedded**.
For example, the resulting indexed **LocalizedProduct** is the following:

```
@Entity
@Table(name = "localized_items")
@Inheritance(strategy = InheritanceType.JOINED)
public abstract class LocalizedItem extends BaseEntity {
    ...
    @Field(termVector = TermVector.YES)
```

```
    @Transient
    public String getText() { return null; }
    ...
}
```

the **TranslatableItem** entity is the following:

```
@Entity
@Table(name = "translatable_items")
@Inheritance(strategy = InheritanceType.JOINED)
public abstract class TranslatableItem extends BaseEntity
{
    ...
    @IndexedEmbedded
    @OneToMany(mappedBy = "translatableItem", cascade = CascadeType.ALL)
    private List<LocalizedItem> localizedItemList;
    ...
}
```

and the resulting **Product** entity is the following:

```
@Indexed
@Entity
@Table(name = "products")
public class Product extends TranslatableItem {
    ...
}
```

6.3 Standard (AND) Search

Search can be performed by creating a query end execute it to retrieve matching entity instances. Query creation can be accomplished by specifying the indexed entity and the fields that should be used for the search. A query can be built as:

```
QueryBuilder qb = fullTextEntityManager.getSearchFactory()
                .buildQueryBuilder().forEntity(Product.class).get();

Query query = qb.simpleQueryString()
                .onFields("localizedItemList.text")
                .matching(matchQuery)
                .createQuery();
```

This will create a query that will search on *localizedItemList* text field for matching keyword specified in *matchQuery* string. This variable is a space-separated list of words. With this query, an entity instance will be retrieved if one of its fields contains at least a keyword contained in *matchQuery*.
If all keywords must be contained in a field to declare the match, the *withAndAsDefaultOperator()* method can be used.
Once the query has been created, the matching entities can be retrieved as:

```
javax.persistence.Query persistenceQuery =
       fullTextEntityManager.createFullTextQuery(query, Product.class);

List<Product> productList = persistenceQuery.getResultList();
```

6.4 Fuzzy Search

Sometimes, keyword exact matching is not the most appropriate way of searching. For example, if a user search for a product and introduces one or more typos in keywords, the correct object could not be retrieved. Even if no typos are introduced, this type of matching could not be satisfying: a user does not know the exact keywords that are included in object's fields.
Fuzzy queries can solve this problem by allowing for approximate matching using the Levenshtein distance algorithm. A fuzzy query can be built as:

```
Query query = qb.keyword()
                .fuzzy()
                .withEditDistanceUpTo(editDistance)
                .withPrefixLength(prefixLength)
                .onFields("localizedItemList.text")
                .matching(matchQuery)
                .createQuery();
```

where *editDistance* specify how much a term can deviate from the other and *prefixLength* define the length of prefix that should be ignore when evaluating the distance.

6.5 Phrase Search

This type of search allows to search for exact or approximate **sentences** in object's fields.

```
Query query = qb.phrase()
                    .withSlop(slop)
                .onField("l"ocalizedItemList.text")
                .sentence(matchQuery)
                .createQuery();
```

where *slop* is the number of other words that can be contained in the specified phrase.

6.6 Similarity Search

Similarity Search can be accomplished through *More Like This* (MLT) queries. This type of query will return all the entities which fields are similar to the one with the requested identifier. The query can be built as:

```
Query query = qb.moreLikeThis()
                .excludeEntityUsedForComparison()
                .favorSignificantTermsWithFactor(2f)
                .comparingField("localizedItemList.text")
                .toEntityWithId(objectId)
                .createQuery();
```

where:

- *excludeEntityUsedForComparison()* can be used to exclude from the returned entities the one used for comparison;
- *favorSignificantTermsWithFactor(float)* can be used to give an higher score to the very similar entities and downgrade the score of mildly similar entities