Deliverable D2.4 – Data Source Identification, Data Requirements and Data Lake Design and Specification Report

Deliverable type	R- Document, Report
Dissemination level	PU - Public
Due date (month)	M13
Delivery submission date	29 February 2024
Work package number	2
Lead beneficiary	IOTIQ Gmbh

This project has received funding from the Horizon Europe Framework Programme of the European Union under grant agreement No. 101094428

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the granting authority can be held responsible for them.

Document Information

Project number	101094428	Acronym	CULTURATI
Project name	Customized Games and Routes For Cultural Heritage and Arts		
Call	HORIZON-CL2-2022-HERITAGE-01		
Topic	HORIZON-CL2-2022-HERITAGE-01-02		
Type of action	HORIZON-RIA		
Project starting date	1 February 2023	Project duration	36 months
Project URL	http://www.CULTURATI.eu		
Document URL	https://culturati.eu/deliverables/		

Deliverable number	2.4			
Deliverable name	Data Source Identification, Data Requirements and Data Lake Design and Specification Report			
Work package number	WP2			
Work package name	System Development and Evaluation			
Date of delivery	Contractual	M13	Actual	M13
Version	Version 1.0			
Lead beneficiary	IOTIQ GMBH			
Responsible author(s)	Metin Tekkalmaz, IOTIQ Gmbh, metin@iotiq.de			
Reviewer(s)	Neşe Şahin Özçelik, Bilkent Universitesi Vakif, nozcelik@bilkent.edu.tr Arzu Sibel İkinci aikinci, Bilkent Universitesi Vakif, aikinci@bilkent.edu.tr Eda Gürel, Bilkent Universitesi Vakif, eda@tourism.bilkent.edu.tr			

Short Description	This deliverable involves identifying and specifying data sources, data requirements, and the design of a data lake. The primary objective is to integrate crowd recognition sensors to enhance the visitor experience through personalized games and routes within indoor and outdoor heritage sites. The deliverable outlines key data sources, a strategic methodology for sensor deployment, data collection protocols, and a data lake design incorporating various database tables. The focus is ensuring accurate, secure, and privacy-compliant data management to facilitate each visitor's engaging and customized digital experience.

History of Changes			
Date	Version	Author	Remarks
21.02.2024	Draft 0.1	Metin Tekkalmaz	First version

26.02.2024	Draft 0.2	Metin Tekkalmaz	Revised after review
08.03.2024	Draft 0.3	Metin Tekkalmaz	Revised fater review
12.03.2024	Final 1.0	Metin Tekkalmaz	Revised fater review

Executive Summary

D2.4 Data Source Identification, Data Requirements, and Data Lake Design and Specification Report marks essential progress in the project CULTURATI. With a focus on data source identification, data requirements, and data lake design, this document outlines a strategic methodology to promote the visitor experience at indoor and outdoor heritage sites through advanced technology integration.

The project emphasizes the transformation of indoor and outdoor heritage sites into interactive spaces. We capture valuable insights into visitor behavior and preferences by incorporating crowd recognition sensors. The success of CULTURATI relies on efficiently harnessing diverse data sources, ensuring accuracy, security, and relevance.

Key achievements include deploying advanced sensors strategically placed within the project's pilot sites in Türkiye as part of STAGE 1 of the project, enabling real-time tracking of visitor movements. This data not only enhances the management of site resources but also forms the basis for personalized and efficient route recommendations, a primary objective of CULTURATI.

The methodology involves a combination of manual input and automatic detection from crowd recognition sensors. Clear data collection protocols and seamless integration with content management systems ensure the accuracy and consistency of information. The system generates efficient routes by correlating crowd data with exhibit content, creating a more engaging and enjoyable experience for each visitor.

The data sources include site databases, map files, authentication servers, content management systems, and crowd recognition sensors. Each source contributes to the comprehensive data requirements, including crowd recognition data, visitor movement data, site-related data, item-related data, route-related data, sensor-related data, content-related data, security and privacy data, access control data, user information, session information, user feedback data, and geospatial data.

The database design incorporates tables such as asked_question, comment, exhibit, exhibition_item, facility, game, game_session, nav_point, sensor, tour_session, user_favorite_exhibition_item, and users, ensuring a structured approach to data storage.

In conclusion, this deliverable lays the groundwork for a seamlessly integrated and technology-driven visitor experience. By identifying, documenting, and optimizing data sources, CULTURATI is committed to transforming cultural and historical preservation through the innovative use of data.

Table of Contents

Executive Summary 4
Table of Contents 6
1. Introduction 8
2. Background 8
3. Data Source Identification 8
3.1. Purpose of Data Source Identification 8
3.1.1. Enhancing Data Quality and Accuracy 8
3.1.2. Enhancing Visitor Experience through Route Optimization 9
3.1.3. Personalised Route Recommendations 9
3.2. Methodology 9
3.2.1. Sensor Deployment Strategy 9
3.2.2. Data Collection Protocols 10
3.2.3. Integration with Content Management (Wiki) System 10
3.3. Types of Data Sources 10
3.4. Data Sources 10
3.4.1. Museum/Site Databases 10
3.4.2. Map files 11
3.4.3. Base Map Data 11
3.4.4. Authentication Server (Keycloak) 12
3.4.5. Content Management System (Wiki) 12
3.4.6. Crowd Recognition Sensors 13
4. Data Requirements 13
4.1. Crowd Recognition Data 13
4.1.1. Visitor Movement Data 13
4.2. Site-related Data 13
4.2.1. Item-related Data 13
4.2.2. Route-related Data 14
4.2.3. Sensor-related Data 14
4.2.4. Sensor Data from Cloud 14
4.3. Content-related Data 15
4.3.1. Content Information 15
4.3.1.1. Category Information 15

4.3.1.2. Level Information 15
4.3.1.3. Tour Related Information 16
4.3.1.4. Game Related Information 16
4.3.2. Metadata 17
4.4. Security and Privacy Data 17
4.4.1. Access Control Data 17
4.4.2. User Information 17
4.4.3. Session Information 17
4.5. User Feedback Data 18
4.5.1. Visitor Feedback 18
4.6. Geospatial Data 18
4.6.1. Site Map 18
4.6.2. Base Map (OpenStreetMap) 18
5. Data Lake Design (Database Design) 20
5.1. asked_question 20
5.2. comment 21
5.3. exhibit 22
5.4. exhibition_item 23
5.5. facility 24
5.6. game 25
5.7. game_session 26
5.8. nav_point 27
5.9. sensor 28
5.10. tour_session 29
5.11. user_favorite_exhibition_item 30
5.12. users 31
5. Conclusion 32

1. Introduction

This document aims to outline the data sources, requirements, and the data lake design involved in CULTURATI, focusing on integrating crowd recognition sensors for personalized visitor experiences. In the evolving landscape of museum/heritage site administration, the effective utilization of data has become pivotal in enhancing visitor experiences, optimizing resource management, and providing a seamless integration of technology within cultural institutions.

2. Background

Traditionally, museums/sites have been repositories of historical and cultural artifacts, but with the advent of technology, the role of museums/sites is expanding to become interactive spaces. Our project aims to revolutionize the visitor experience at venue-based Cultural and Creative Industries (CCIs) including museums/sites by integrating a sophisticated system that not only captures and manages data related to exhibits but also utilises sensors for crowd recognition. This allows administrators to gain valuable insights into visitor behavior and preferences as well as help visitors spend their time more efficiently.

3. Data Source Identification

The success of CULTURATI hinges on the ability to harness and leverage diverse data sources effectively. Identifying these data sources is the cornerstone of creating a robust foundation for our system. Understanding where our data originates is fundamental to ensuring its accuracy, security, and relevance. The insights derived from these data sources will not only enhance the management of site resources but also contribute to a personalized and engaging experience for our visitors.

3.1. Purpose of Data Source Identification

Below are the purposes of data source identification:

- To ensure efficient management of museum/site data.
- To enhance the visitor experience by utilizing crowd recognition data for personalized content / better route recommendations.
- To maintain data quality, security, and compliance.

3.1.1. Enhancing Data Quality and Accuracy

Accurate and reliable data is paramount to the success of CULTURATI. The purpose of identifying data sources is to ensure the integrity and quality of the information used within the system. By

clearly understanding the origin of data, we can implement measures to validate, clean, and maintain the accuracy of the content related to exhibits and visitor preferences.

3.1.2. Enhancing Visitor Experience through Route Optimization

One of the primary objectives of CULTURATI is to significantly enhance the visitor experience by providing personalized and efficient routes through the site. To achieve this goal, the system relies on analyzing crowd data gathered through advanced sensors strategically placed within the site.

3.1.3. Personalised Route Recommendations

CULTURATI aims to provide a personalized experience for visitors. Identifying data sources allows us to understand visitor preferences, interests, and behaviors. This information is then used to recommend tailored routes and content, creating a more engaging and enjoyable experience for each visitor. In this regard, the purpose of data source identification in this context is twofold;

- Accurate Crowd Data Collection: Identifying the sources of crowd data allows us to ensure the accuracy and reliability of the information collected. The integration of sensors enables real-time tracking of visitor movements, providing valuable insights into crowd density, flow patterns, and popular exhibit areas.
- Optimization of Route Recommendation Algorithm: We can refine and optimize the algorithms responsible for generating route recommendations by identifying data sources. We can create more efficient visitor routes by correlating crowd data with content-related information about items/objects on site.

3.2. Methodology

Data source identification will be conducted through a combination of manual input from administrators and automatic detection from crowd recognition sensors. Utilization of software tools designed explicitly for managing site data and integrating sensor-generated data.

3.2.1. Sensor Deployment Strategy

The first step in our methodology involves strategically deploying sensors throughout the premises/heritage site. This process is designed to capture comprehensive crowd data, including visitor movements, dwell times, and popular exhibit areas. The deployment strategy considers factors such as exhibit layouts, visitor traffic patterns, and potential congestion areas.

3.2.2. Data Collection Protocols

Establishing clear data collection protocols is crucial for ensuring the accuracy and consistency of the gathered information. These protocols define the frequency of data capture, the data types to be collected (e.g., location data, visitor interactions), and the methods used to anonymize and secure sensitive information. By adhering to these protocols, we maintain the integrity of the data and uphold privacy standards.

3.2.3. Integration with Content Management (Wiki) System

Integrating with the content management system is a key aspect of our methodology. This involves mapping the data sources related to exhibit content, which includes item descriptions, historical information, and multimedia content. By establishing a connection between crowd data and exhibit content, we create a foundation for generating personalized route recommendations based on visitor interests and crowd data.

3.3. Types of Data Sources

- Site Database: Contains information about the locations and/or general information about exhibits, artifacts, and sensors provided by administrators.
- Map files: Geospatial information of indoor/outdoor elements of the site.
- Authentication Server: Contains information about users and their sessions.
- Content Management System(wiki) Database: Contains content information(questions, information pieces etc.) related to the exhibition items.
- Crowd Recognition Sensors: Capture and analyze real-time data on visitor movement and preferences.

3.4. Data Sources

3.4.1. Museum/Site Databases

- Source Name: SiteDB
- Description: Central repository for exhibition item, exhibit, and sensor information.
- Data Format: Structured
- Frequency of Updates: Regular updates based on new exhibits or location changes.
- Data Owner: Museum/Site Administrator
- Access Controls: Role-based access controls for administrators.

There is one museum/site database (relational database) for each site which is utilised by the CULTURATI core application, which is a multi-tenant web application (database per tenant). This approach is selected to achieve the highest level of data isolation between tenants.

Users' personal information will not be stored in this database since it is stored on the authentication server. In cases this information is fetched from the authentication server, they will not be cached or stored in the local storage.

3.4.2. Map files

- Source Name: MapFiles
- Description: Json based map file of the museum/site.
- Data Format: Json (GeoJson)
- Frequency of Updates: Rare, since it is based on the floor plan.
- Data Owner: Museum/Site Administrator
- Access Controls: Role-based access controls for administrators.

The map file of the museum/site is typically the floor plan/plan information which is stored in Json format. The file is supposed to be generated using JOSM (https://josm.openstreetmap.de/) or similar software and imported into CULTURATI using the core application. No personal or commercially confidential information will be stored in these files.

3.4.3. Base Map Data

- Source Name: BaseMap
- Description: OpenStreetMap is a free, open geographic database updated and maintained by a community of volunteers via open collaboration. Contributors collect data from surveys, trace from aerial imagery and also import from other freely licensed geodata sources. The data is fetched as separate “tiles”, those are the most basic part of the base map data. Web maps work by loading map tiles from a tile server. The tiles can differ according to the zoom levels and the coordinates.
  (https://wiki.openstreetmap.org/wiki/Slippy_map_tilenames)
  - Tiles are typically 256 × 256 pixel PNG files
  - Each zoom level is a directory, each column is a subdirectory, and each tile in that column is a file
  - Filename(url) format is /zoom/x/y.png
- Data Format: png

- Frequency of Updates: Continuous.
- Data Owner: OpenStreetMap
- Access Controls: none

OpenStreetMap functions as a freely accessible geographic database, constantly refreshed and managed by a collaborative community of volunteers. These contributors acquire information through surveys, trace routes from aerial images, and import data from other openly licensed geodata sources. The gathered data is retrieved in the form of distinct "tiles," serving as fundamental components of the base map data. Web maps operate by retrieving these map tiles from a dedicated tile server, with variations based on zoom levels and coordinates..

3.4.4. Authentication Server (Keycloak)

- Source Name: AuthServer
- Description: Authentication Server used to register/authenticate the users into the application. Keycloak stores data related to the users and their sessions.
- Frequency of Updates: Frequent updates of the user data.
- Data Owner: Museum/Site Administrator / CULTURATI Administrator
- Access Controls: Restricted access to authorized personnel.

Keycloak allows single sign-on with identity and access management in the project. There will be only one instance of Keycloak throughout the project in which different realms will be used to manage the content creator and data entry operator accounts for each site. While registering, visitors will not be asked for personal information like name, surname, or phone number. It's the administrator's decision to store this information for administrative users like content creators or data entry operators.

3.4.5. Content Management System (Wiki)

- Source Name: SiteDB
- Description: Central repository for exhibition item related content.
- Data Format: Structured
- Frequency of Updates: Regular updates based on edited content.
- Data Owner: Museum/Site Administrator / Content Creators
- Access Controls: Role-based access controls for administrators.

One content management system instance per site will be deployed to store each site's content-related information. The content will be stored in a relational database.

3.4.6. Crowd Recognition Sensors

- Source Name: CrowdSensors
- Description: Real-time data on crowd density.
- Data Format: Json
- Frequency of Updates: On-Demand / Real-time
- Data Owner: Museum/Site IT Department
- Access Controls: Restricted access to authorized personnel.

Crowd sensors are installed on the sites, and their data is collected by the sensor provider on the cloud. This data is being collected by the AI component of CULTURATI using the provider APIs. No personal or commercially confidential information will be fetched from the sensors.

4. Data Requirements

4.1. Crowd Recognition Data

4.1.1. Visitor Movement Data

- Source: CrowdSensors
- Description: Density of visitors in different areas of the Museum/Site.
- Attributes:
    - Timestamp
    - Area or zone identifier
    - Number of visitors within the defined space

4.2. Site-related Data

4.2.1. Item-related Data

- Source: SiteDB
- Description: Information related to the exhibition items.
- Attributes:
    - Item ID
    - Title
    - Wiki ID

- Location information

4.2.2. Route-related Data

- Source: SiteDB
- Description: Information related to the footways within the site.
- Attributes:
  - Navigation Point ID
  - Name
  - Description
  - Location information

4.2.3. Sensor-related Data

- Source: SiteDB
- Description: Information related to the sensors.
- Attributes:
  - Sensor ID
  - Type
  - Name
  - Description
  - Location information

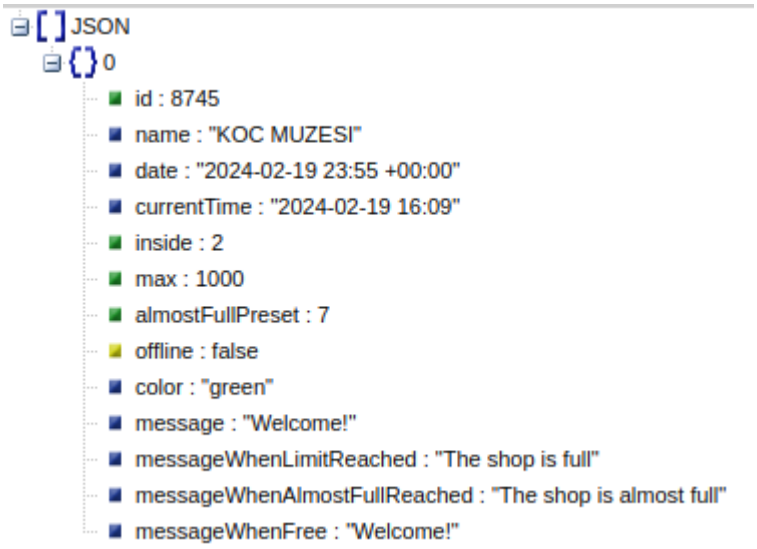
4.2.4. Sensor Data from Cloud

- SensMax APIs
- Description: Crowd data from the sensors.
- Attributes:
  - Sensor ID
  - Name
  - Date
  - Time
  - Inside (people count)
  - Max
  - AlmostFullpreset
  - Offline
  - Color

- Message
- MessageWhenLimitReached
- MessageWhenAlmostFullReached
- MessageWhenFree

Getting Crowd Data from Sensor Provider (SensMax):

- Crowd data is fetched from the sensors in the following format using the SensMax cloud enpoints (https://my.sensmax.eu/api/v2/report/<report_id>) :



Here, the “inside” field is important for deciding how crowded this area is, since it includes the count of the people in the area.

4.3. Content-related Data

4.3.1. Content Information

4.3.1.1. Category Information

- Source: WikiDB
- Description: Category related information in the Museum/Site.
- Attributes:
  - Category ID
  - Content language
  - Name

4.3.1.2. Level Information

- Source: WikiDB
- Description: Level related information in the Museum/Site.

- Attributes:
    - Level ID
    - Content language
    - Name

4.3.1.3. Tour Related Information

- Source: WikiDB
- Description: Tour related information about the items in the Museum/Site.
- Attributes:
    - Content ID
    - Content language
    - Category
    - Level
    - RelatedItem
    - Title
    - InfoPiece
    - Multimedia content (images, videos)
    - WebsiteURL

4.3.1.4. Game Related Information

- Source: WikiDB
- Description: Game related information (questions) about the items in the Museum/Site.
- Attributes:
    - Question ID
    - Content language
    - Category
    - Level
    - RelatedItem
    - QuestionBody
    - AsnwerType
    - Options
    - Hint
    - Points

- Answer
- GameType
- Multimedia content (images, videos)
- WebsiteURL

4.3.2. Metadata

- Source: WikiDB
- Description: Additional metadata associated with each exhibit for categorization and recommendation purposes.
- Attributes:
  - Content ID
  - Keywords or tags (Category/Level)

4.4. Security and Privacy Data

4.4.1. Access Control Data

- Source: AuthServer
- Description: Information about who has access to the collected data and under what conditions.
- Attributes:
  - User ID or role
  - Data access permissions

4.4.2. User Information

- Source: AuthServer
- Description: Information about the user.
- Attributes:
  - User ID
  - Role
  - Email

4.4.3. Session Information

- Source: AuthServer
- Description: Information about the user sessions.
- Attributes:

- User ID
- Session ID
- Time data (Start/End)

4.5. User Feedback Data

4.5.1. Visitor Feedback

- Source: SiteDB
- Description: Feedback provided by visitors regarding the recommended routes and overall Museum/Site experience.
- Attributes:
  - Timestamp
  - Visitor ID or pseudonymous identifier
  - Feedback comments or ratings

4.6. Geospatial Data

4.6.1. Site Map

- Source: MapFile
- Description: Json file containing the geospatial information of the entities that form the plan for the site. (or floor plan).
- Attributes:
  - Properties
    - Indoor / Outdoor
    - Level
    - Type (Footway, concrete, etc.)
  - Geometry
    - Type
    - Nodes
    - Coordinates

4.6.2. Base Map (OpenStreetMap)

- Source: BaseMap
  - Description: OpenStreetMap tiles coming from the tile server.
- Attributes:

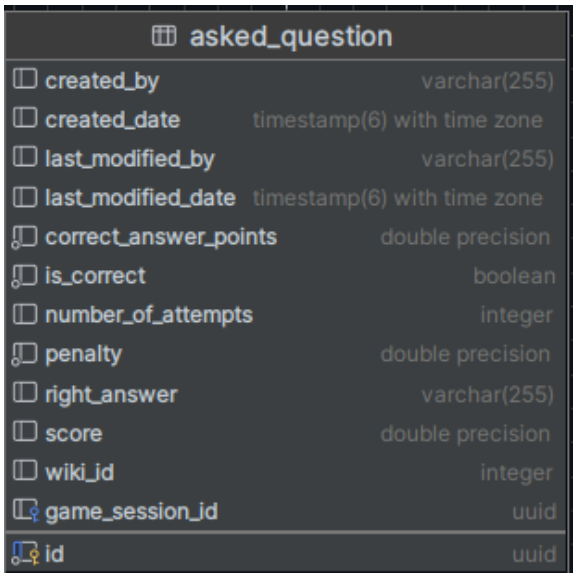
- Properties
  - Zoom Level (https://wiki.openstreetmap.org/wiki/Zoom_levels)
  - X (Calculated according to the longitude)
  - Y (Calculated according to the lattitude)

5. Data Lake Design (Database Design)

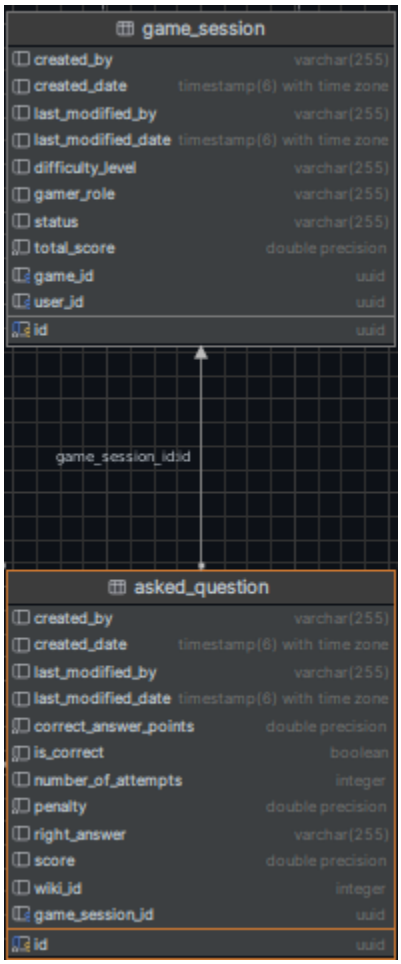
The database is designed to have the following tables (in alphabetical order)

5.1. asked_question

Stores the information about the question asked to the users.

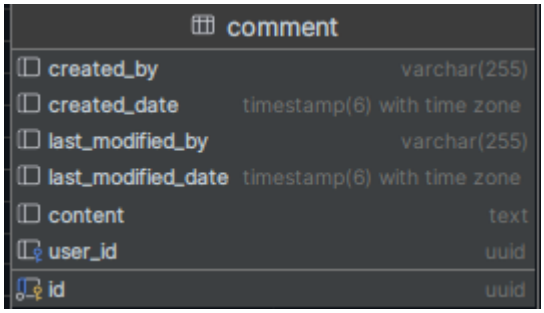


Relations:

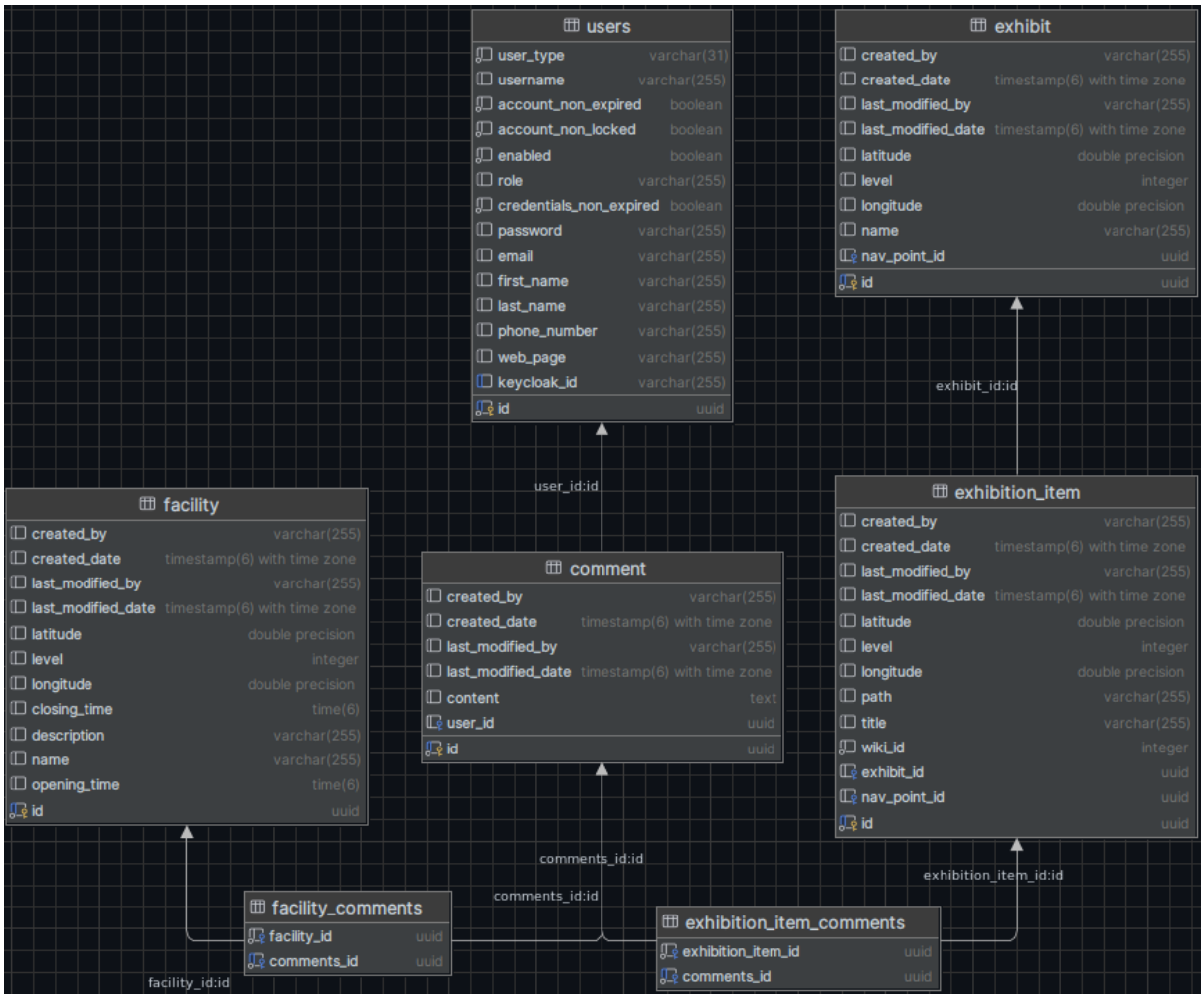


5.2. comment

Stores the comments of the application users about the items/facilities.

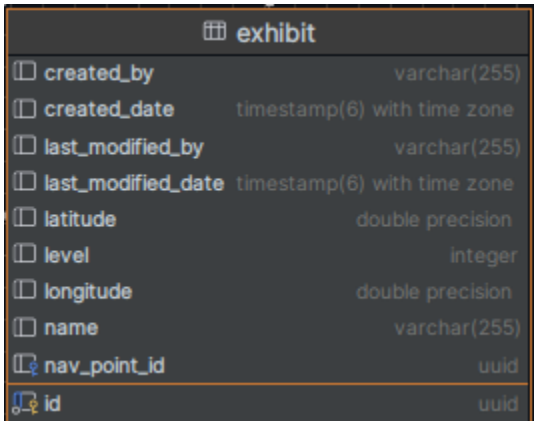


Relations:

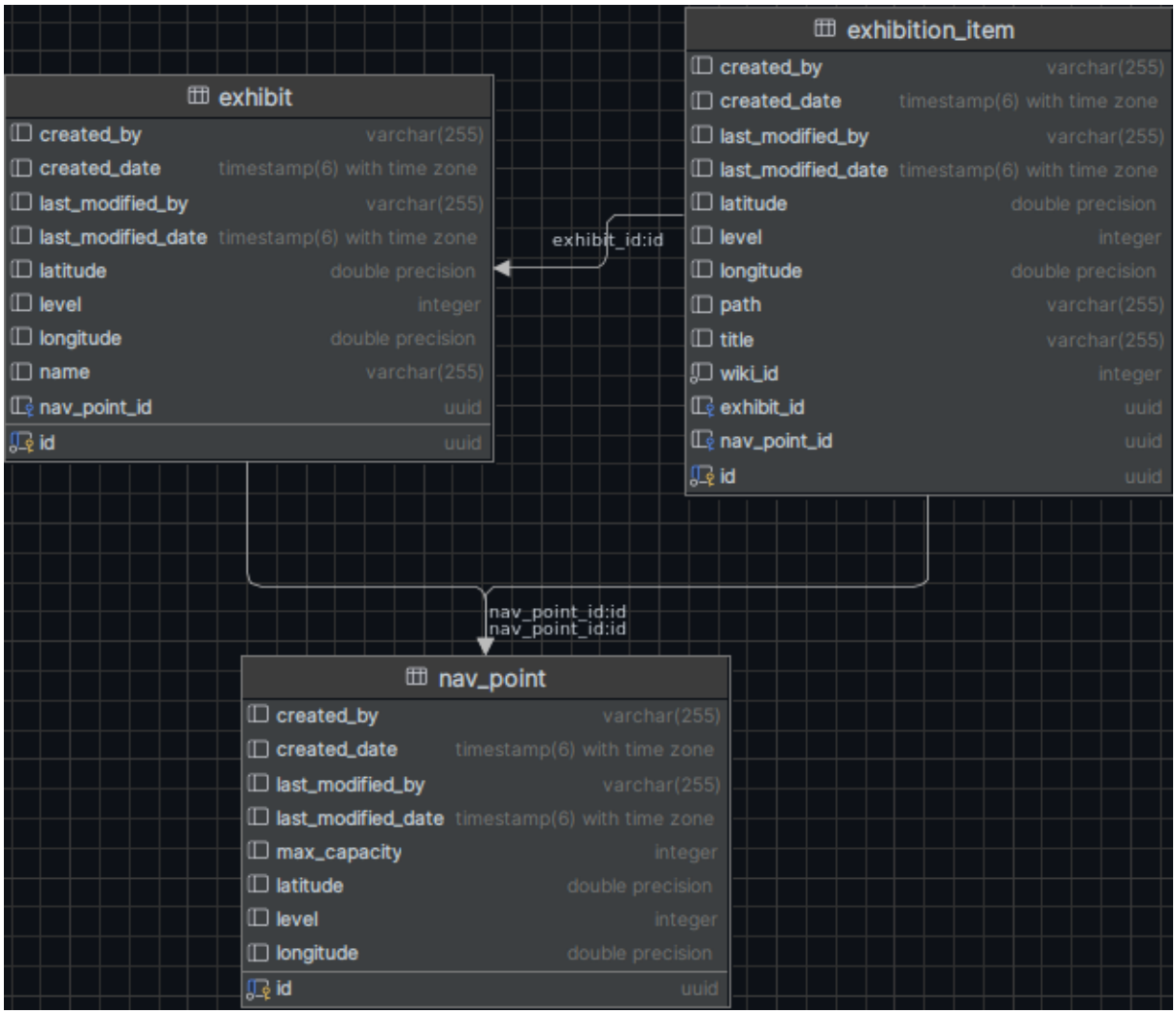


5.3. exhibit

Stores the information related to the exhibits on the site.

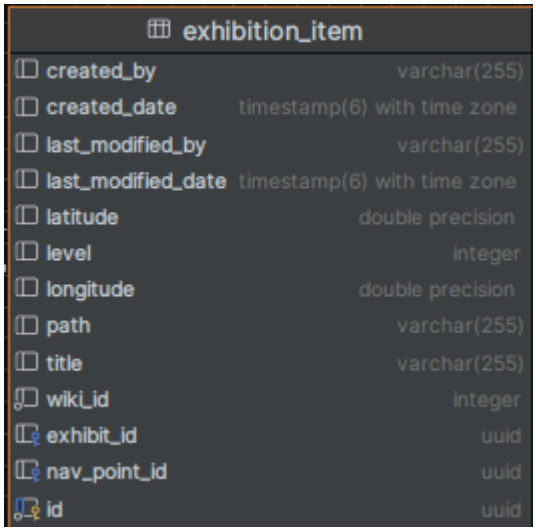


Relations:

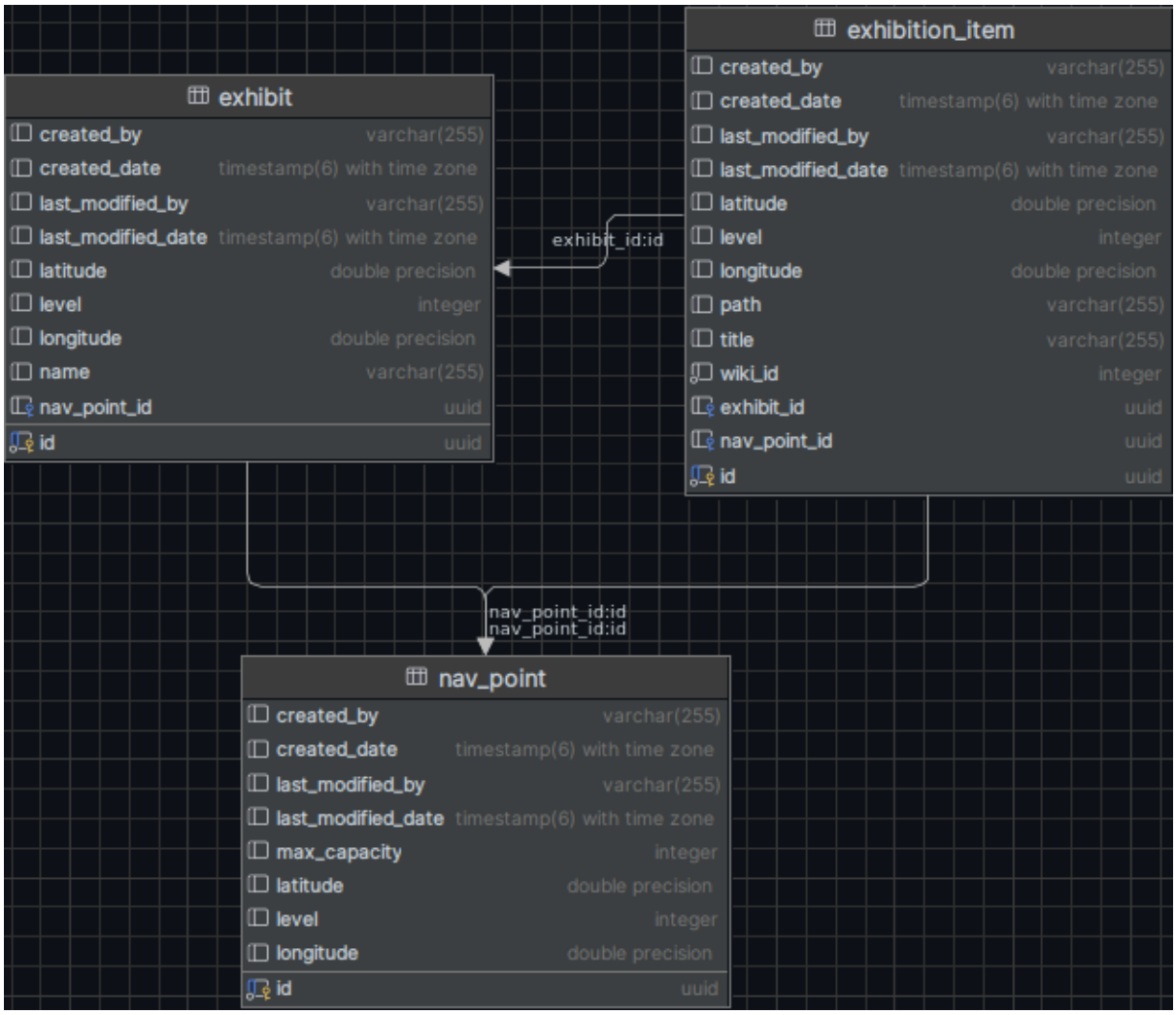


5.4. exhibition_item

Stores the information related to the exhibition items on the site.

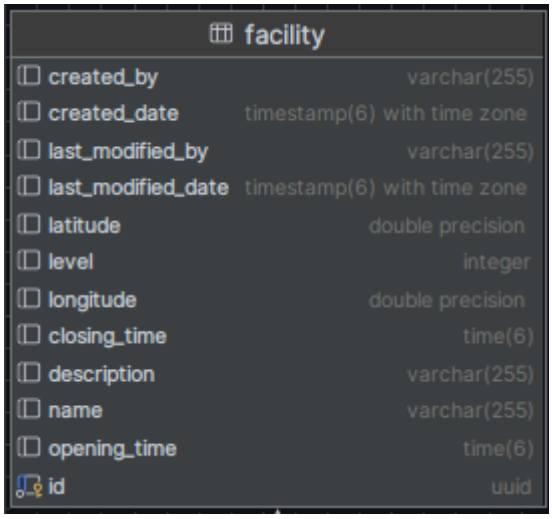


Relations:

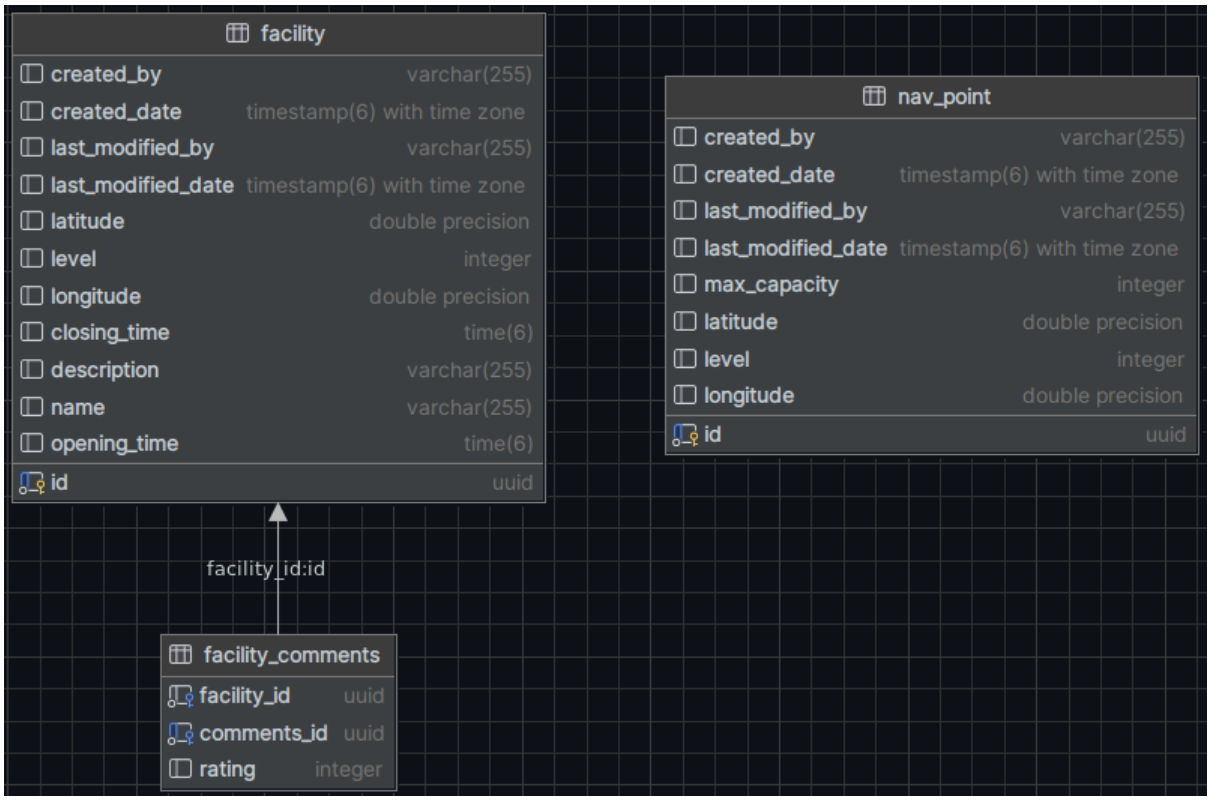


5.5. facility

Stores the information related to the facilities on the site.



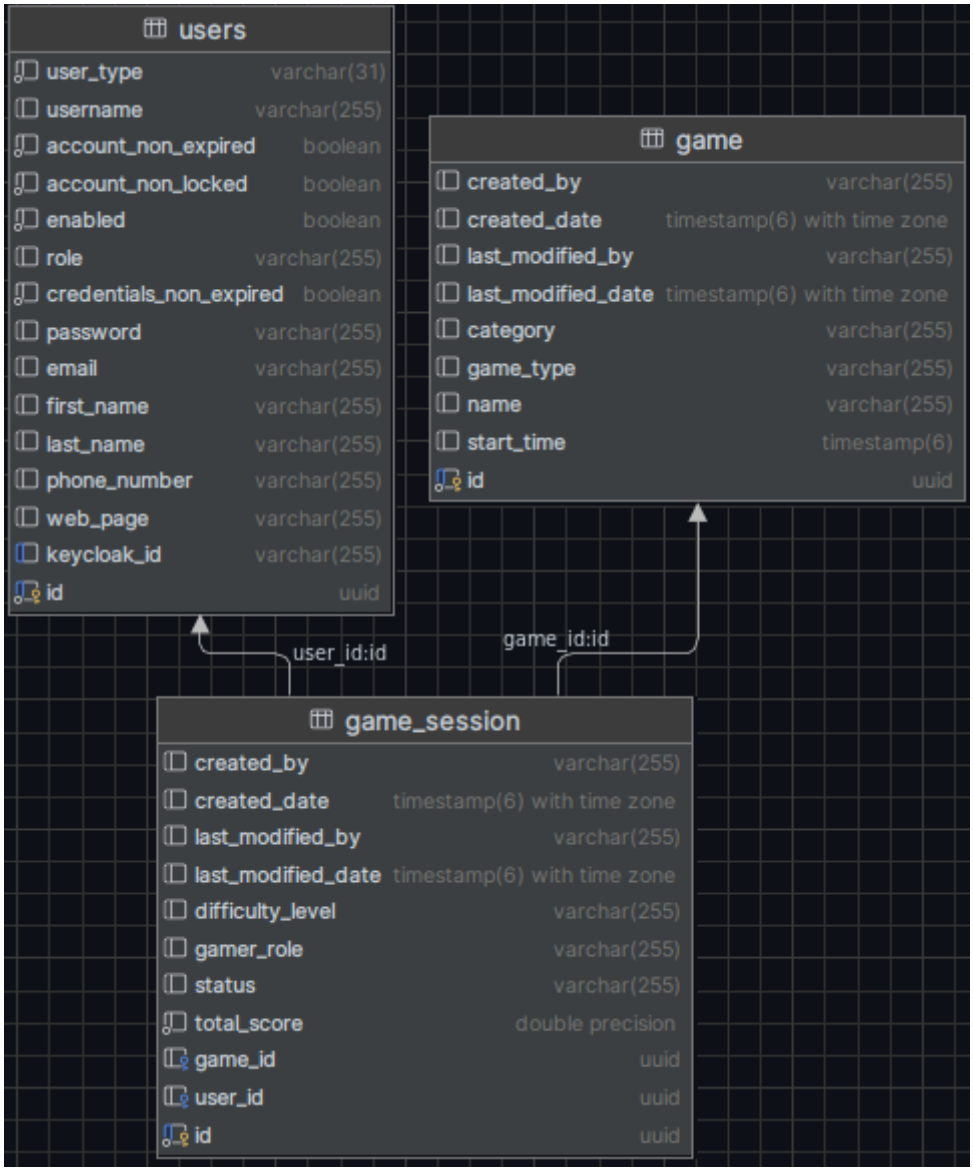
Relations:



5.6. game

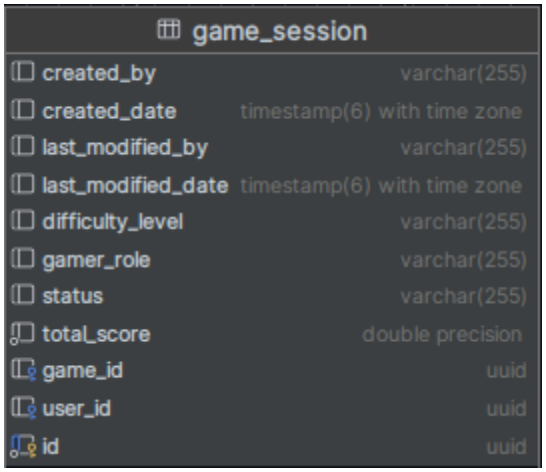
Stores the information related to the games created by the visitors.

Relations:

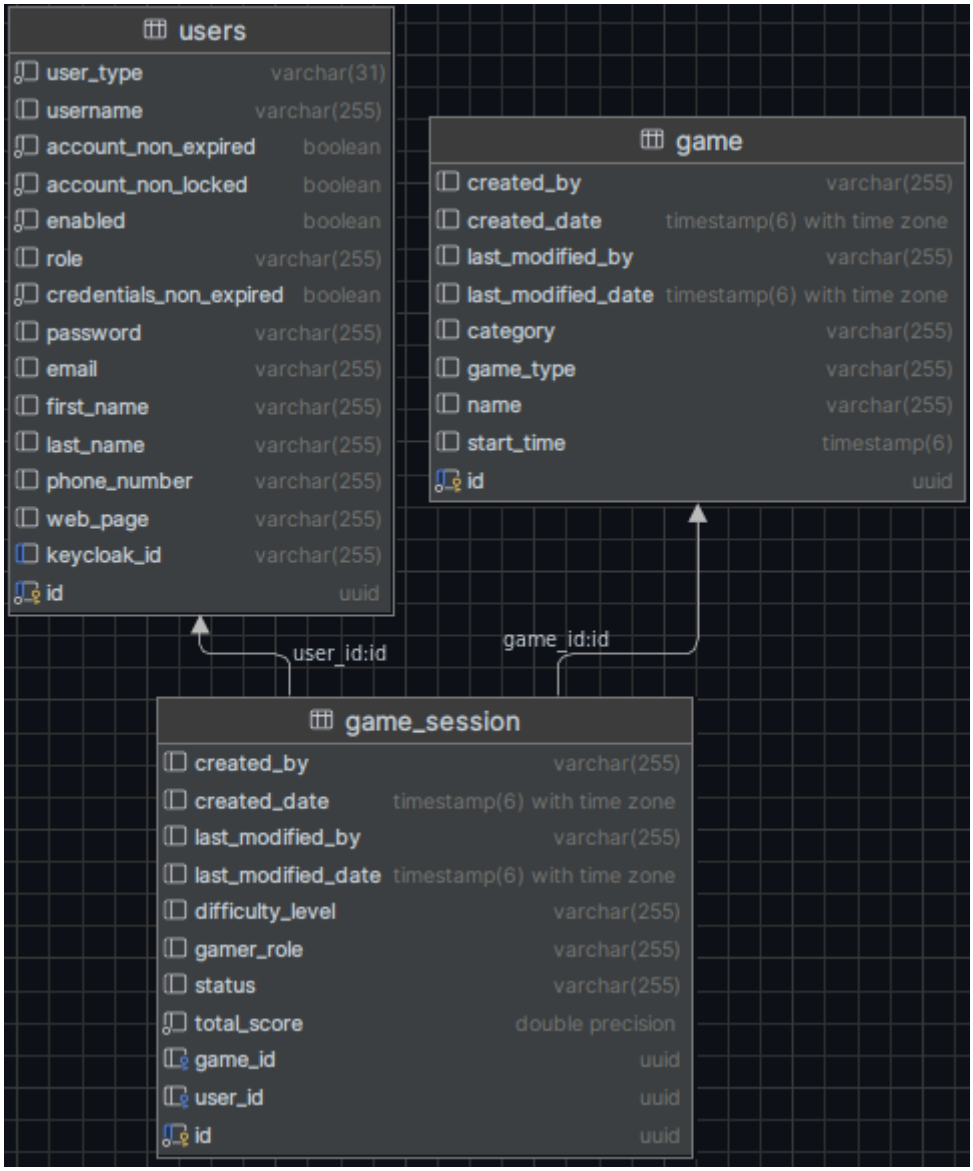


5.7. game_session

Stores the information related to the user sessions in the games.

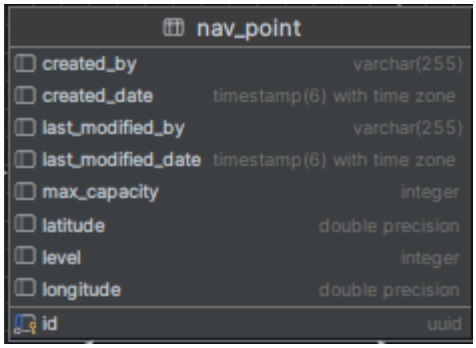


Relations:

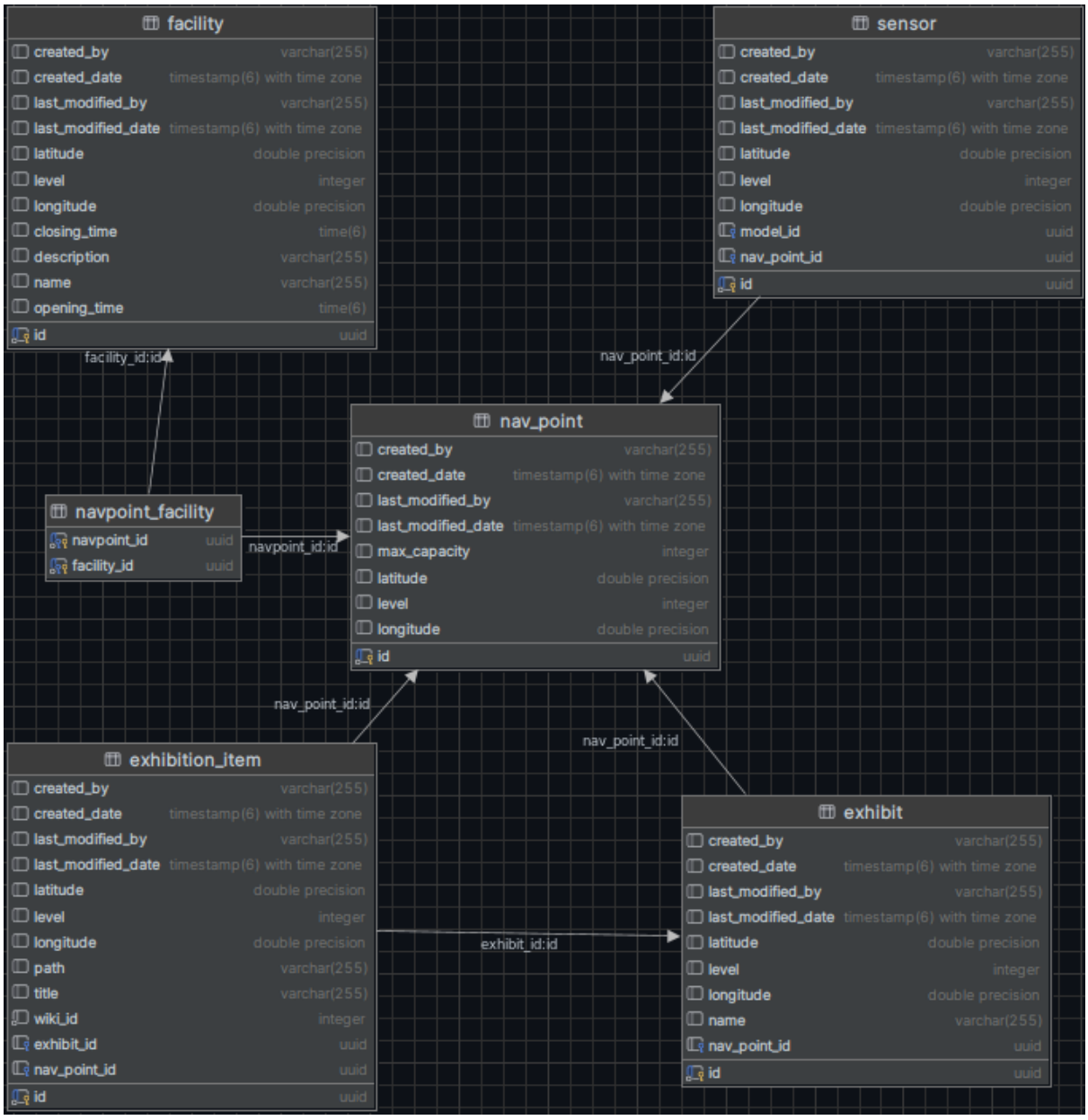


5.8. nav_point

Stores the information related to the navigation points that are used for navigating the users.

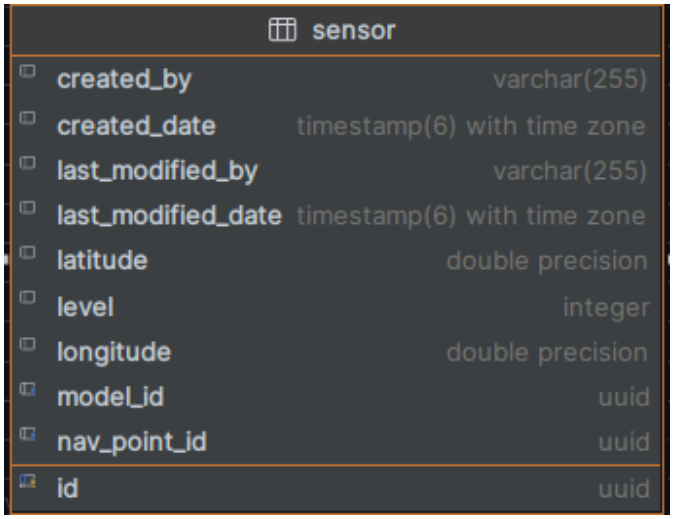


Relations:

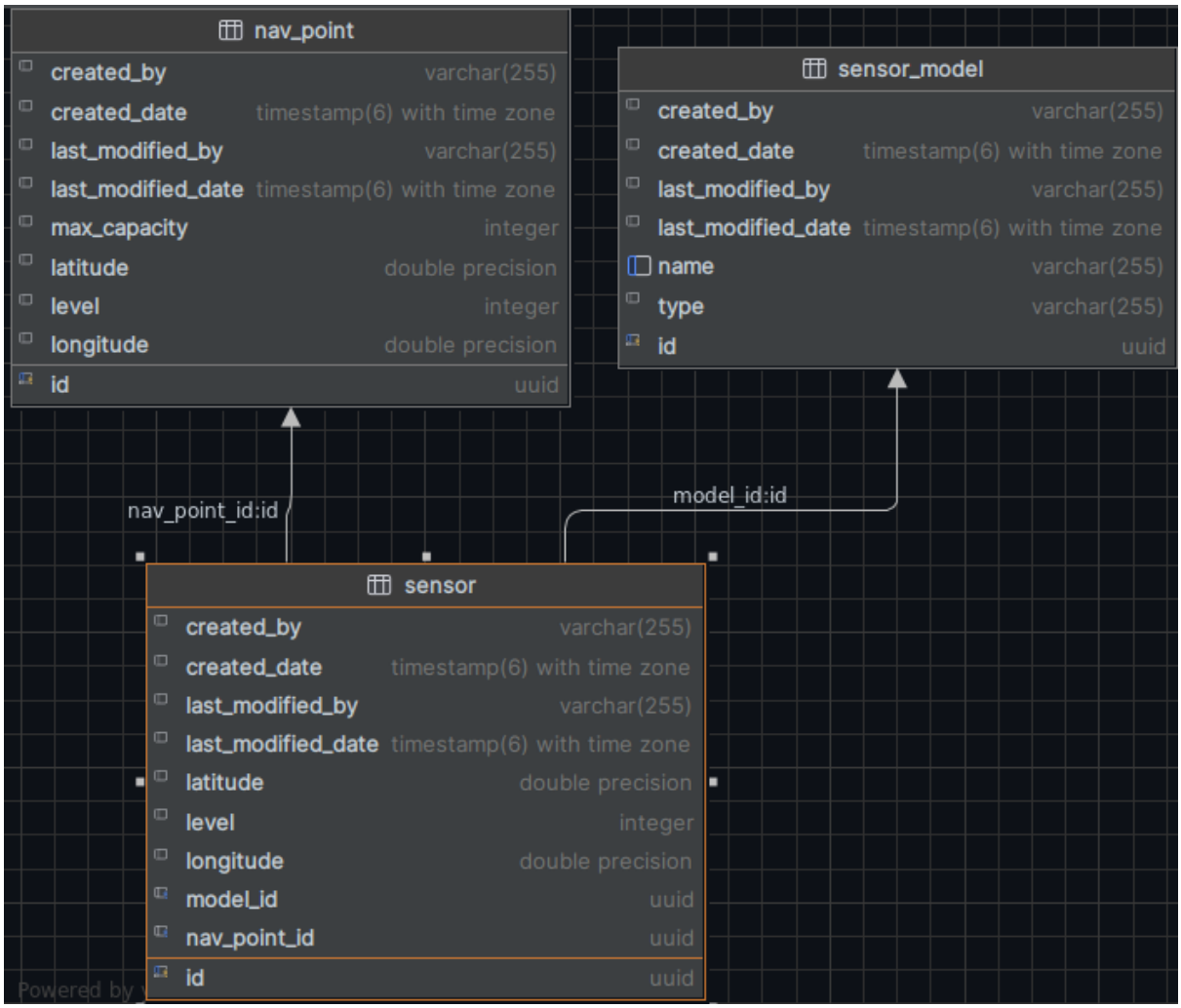


5.9. sensor

Stores the information related to the sensors in the site.

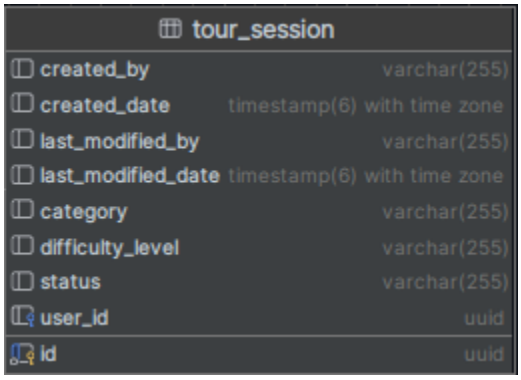


Relations:

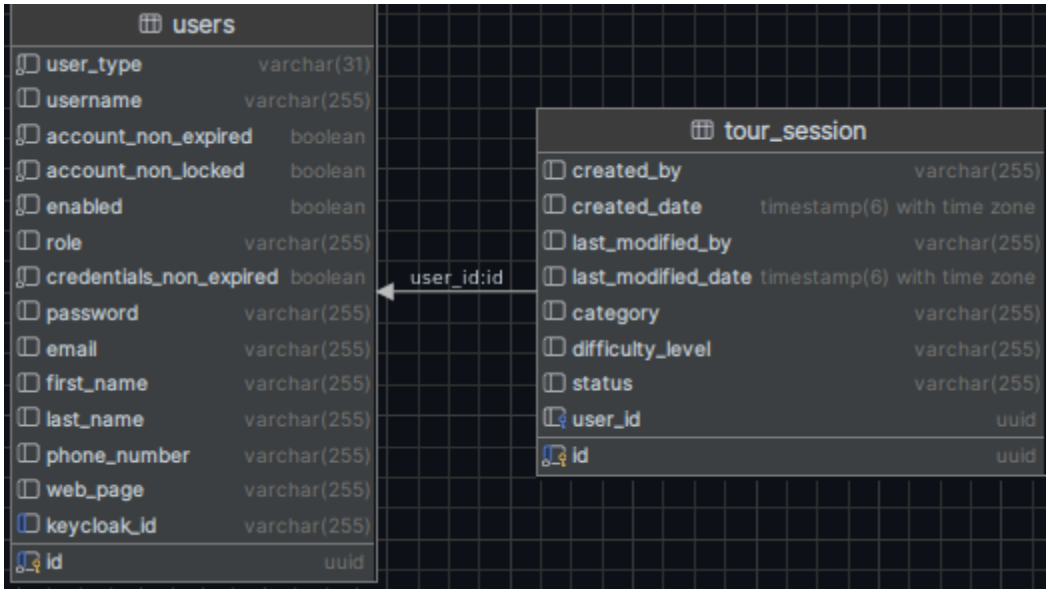


5.10. tour_session

Stores the information related to the tours created by the ai component for the visitors.

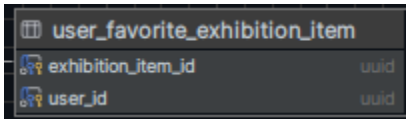


Relations:

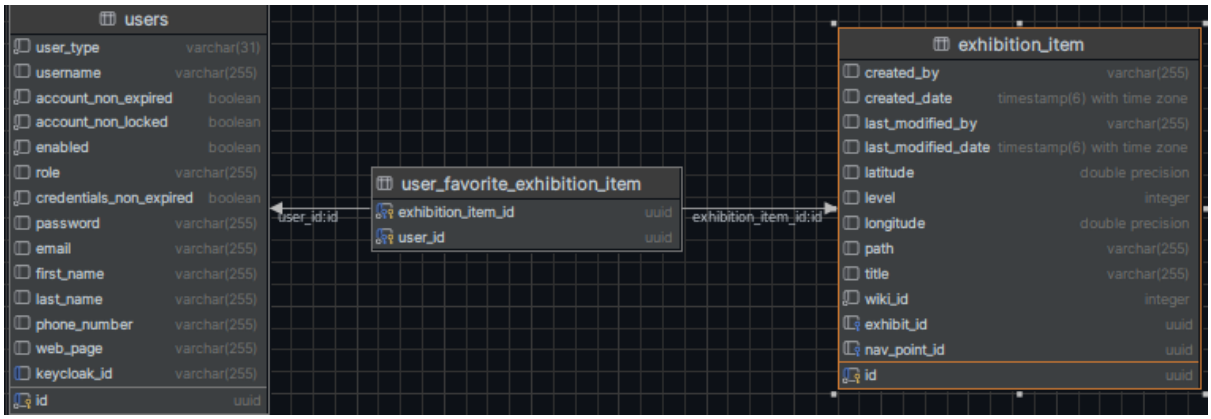


5.11. user_favorite_exhibition_item

Stores the information related to the visitor's favorite exhibition items.

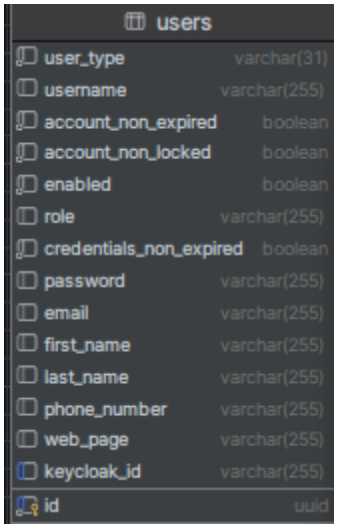


Relations:

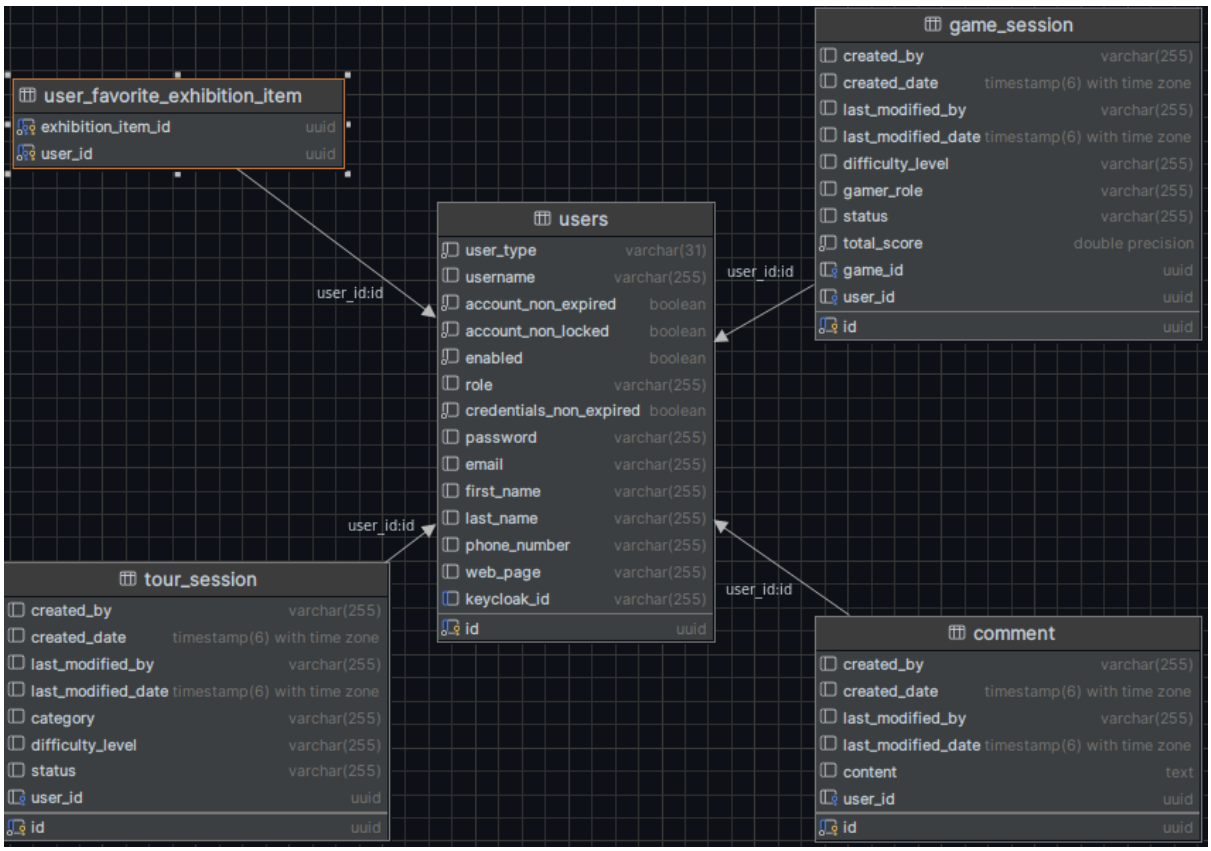


5.12. users

Stores the information related to the users of the system.



Relations:



5. Conclusion

In the pursuit of creating an innovative system like CULTURATI, the comprehensive identification and understanding of data sources play a pivotal role. This document has outlined a strategic methodology for recognizing and documenting diverse data sources, specifically focusing on crowd recognition for route optimization within our sites.

The integration of advanced sensors capturing crowd data, coupled with a robust content-related database, forms the backbone of our system. Through diligent identification and documentation, we ensure the accuracy, security, and relevance of the data driving our Site Management System.

The methodology outlined in this document provides a structured approach to deploying sensors strategically, defining data collection protocols, and integrating with content management systems. Continuous monitoring, stakeholder collaboration, and adherence to security and compliance standards further underscore our commitment to excellence in data management.

In conclusion, by identifying, documenting, and continuously optimizing our data sources, we lay the groundwork for a system that not only meets the expectations of administrators but also delights and engages visitors on a profound level. The journey towards a seamlessly integrated and technology-driven visitor experience starts with the meticulous identification of our data sources, and this document serves as our guide in that pursuit.

As we move forward, we are committed to the principles of accuracy, privacy, and innovation, ensuring CULTURATI stands as a testament to the transformative potential of leveraging data in cultural and historical preservation.