Contenido

[Group Hands-on Assignment 1Tarea 2](#_Toc118882028)

[Deliverables 2](#_Toc118882029)

[Group Hands-on Assignment 2Tarea 3](#_Toc118882030)

[Deliverables 3](#_Toc118882031)

[Group Hands-on Assignment 3Tarea 4](#_Toc118882032)

[Deliverables 4](#_Toc118882033)

[Group Hands-on Assignment 4Tarea 4](#_Toc118882034)

[Deliverables 5](#_Toc118882035)

[Group Hands-on Assignment 5Tarea 5](#_Toc118882036)

[Deliverables 5](#_Toc118882037)

[Group Hands-on AssignmentTarea 6](#_Toc118882038)

[Group Hands-on Assignment 7 (optional)Tarea 6](#_Toc118882039)

[Deliverables 7](#_Toc118882040)

# [Group Hands-on Assignment 1Tarea](https://moodle.upm.es/titulaciones/oficiales/mod/assign/view.php?id=47001)

The goal of this assignment is to find some datasets that can be used by the students in different assignments during the rest of the course and to specify the application that they want to build on top of the datasets.

The selected datasets will be the one that the students will transform into RDF, link with other datasets, and publish online.

The students have to search in the Web for a dataset that satisfies the following requirements:

* R1. Data are in the smart cities domain.
* R2. Data are available as a CSV file.
* R3. Data have an open license so it can be published.
* R4. Data can be easily linked with generic real-world entities (e.g., locations).
* R5. Documentation exists for the data (optional).
* R6. Data can come from multiple data sources (optional).

## Deliverables

Each group will create a directory in github for uploading all the materials of the hands-on. Group directories must follow the format "GroupXX" where "XX" is the number of the group (e.g., "Group01"). Group numbers will be created in a first come-first served basis, starting from number 1 and giving consecutive numbers.

The deliverables required for this assignment are:

* A Markdown document with the names and the GitHub users of the members of the group ("README.md" in the root of the group directory). ok
* CSV files with the datasets selected (under a "csv" directory).
* An HTML document that discusses whether the selected datasets satisfy the requirements. Requirements R1 to R4 must be satisfied by the datasets with the exception of the optional requirements (R5-R6); for these, the degree of satisfaction must be discussed ("datasetRequirements.html" under a "requirements" directory).

Utilizamos el dataset 200342-0-centros-dia.csv que recoge los centros de día de la ciudad de Madrid indicando su ubicación, servicios, accesibilidad y forma de acceso por transporte público. Cumple con los siguientes requisitos:

* R1. Data are in one of the selected smart city domains. Corresponde a servicios dentro de Smart community y Smart community facilities. Incluiso, desde un punto de vista sociosanitario, podría incluso vincularse con Smart medical treatment como lugar para paliar la soledad
* R2. Data are available as a CSV file. Sí, se pueden descargar desde el portal de datos abiertos del ayuntamiento de Madrid. Enlace <https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=22bceca8a5a03410VgnVCM1000000b205a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default> . Se pueden descargar como csv, json, rdf, xml o acceder a ellos vía API.
* R3. Data have an open license so they can be published. No se especifica la licencia de los datos pero en el enlace “Publicación de datos” cuando se hace referencia a “Licencia” sí se indica expresamente que “permiten la reutilización de los documentos y datos sometidos a ellas para fines comerciales y no comerciales.” además de incluir ciertas condiciones generales para la reutilización que sumado a su libre uso indica que corrsponde a una open license.
* R4. Data can be easily linked with generic real-world entities (e.g., locations). Sí, se puede enlazar con locations, bus stop, paradas de metro, barrios y distritos.
* R5. Documentation exists for the data (optional). Sí, dentro de la documentación asociada, existe una Estructura del conjunto de datos, donde se detallan los datos.
* R6. Data can come from multiple data sources (optional). No, la única fuente de datos es el archivo 200342-0-centros-dia.csv provisto por el portal de datos abiertos del Ayuntamiento.

We use *200342-0-centros-dia.csv* dataset which includes all the day care centres in Madrid, indicating its location, services, access and how to reach the place using public transport. It comply with the following requirements:

* R1. Data are in one of the selected smart city domains. It corresponds to Smart community and the Smart community facilities services. Even, from the socio sanitary point of view, it could be linked with Smart medical treatment as a place to cure loneliness.
* R2. Data are available as a CSV file. It is posible to download it from the [Madrid City Council Open Data Site.](https://datos.madrid.es/egob/catalogo/200342-0-centros-dia.csv) It can also be downloaded as a csv, json, rdf, xml or via API
* R3. Data have an open license so they can be published. The site doesn´t especify a specific license for the data. However, in the link “Publicación de datos” when there is a reference to ”[Licencia](https://datos.madrid.es/egob/catalogo/aviso-legal)” it is indicated that “permiten la reutilización de los documentos y datos sometidos a ellas para fines comerciales y no comerciales”, furthermore it mentions some general conditions for data freely reused and redistributed by anyone which corresponds to an open license.
* R4. Data can be easily linked with generic real-world entities (e.g., locations). The data can be linked with locations, bus stops, metro stations, neibourghoods and districts
* R5. Documentation exists for the data (optional). Yes, among the associated documentation exists an link to “[Estructura del conjunto de datos](https://datos.madrid.es/FWProjects/egob/Catalogo/XComun/Ficheros/Estructura_DS_ConjuntoDatos.pdf)” where the data is explained.
* R6 – The only data source available is the file *200342-0-centros-dia.csv* which is provided by the [Madrid City Council Open Data Site](https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=22bceca8a5a03410VgnVCM1000000b205a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default).
* An HTML document that specifies the concrete requirements of the application that will be built on top of the selected datasets and that included mock-ups of the expected user interfaces ("applicationRequirements.html" under a "requirements" directory).
* A Markdown document with the hands-on self-assessment ("selfAssessmentHandsOn1.md" in the root of the group directory).

# [Group Hands-on Assignment 2Tarea](https://moodle.upm.es/titulaciones/oficiales/mod/assign/view.php?id=47010)

The goal of this assignment is to get familiar with the first steps in the Linked Data generation process.

The students will have to take their selected dataset(s) and perform the following tasks:

* **Analyse Data Set**. Analyse both the data (quantities, value ranges, etc.) as well as the schema of the data.
* **Analyse Licensing of the Data Source**. Analyse who is the publisher of the dataset and its rightsholder, as well as the licence of the dataset. Define the potential license to be used for the dataset to be generated.
* **Define Resource Naming Strategy**. Define the resource naming strategy to be followed for the ontology and the data to be generated (URI form, content negotiation, URIs domain, path, patterns, etc.).
* **Develop Ontology**. Develop a lightweight ontology that includes the classes and properties (with their domains and ranges) to be used with the dataset. The rest of ontological components (e.g., restrictions) are optional for this assignment.

**Assignment materials**

This assignment has to be performed with the datasets selected by each group of students.

## Deliverables

The deliverables must be uploaded to the hands-on group directory in github (i.e., "/HandsOn/GroupXX/").

The deliverables required for this assignment are:

* An HTML document that includes the analyses performed over the data source and the licensing of the data source, as well as the resource naming strategy defined ("analysis.html" in the root of the group directory).
* An OWL file with the ontology developed, according to the resource naming strategy defined, and using the Turtle syntax ("\*.ttl" under an "ontology" directory).
* An RDF file with a sample instantiation of the ontology, following the resource naming strategy, and using the Turtle syntax (“\*-example.ttl” under an "ontology" directory).
* A Markdown document with the hands-on self-assessment (”selfAssessmentHandsOn2.md" in the root of the group directory).

Dudas:

* ¿hemos hecho el data characteristics y shcema? Parece que nos dice que lo hagamos con open refine.
* ResourceName Strategy no tengo claro si / o # [diapositiva 35](https://moodle.upm.es/titulaciones/oficiales/pluginfile.php/10225623/mod_resource/content/9/ODKG%2007.MGLD.pdf) Había / y parece que vale.
  + El criterio que hemos puesto y no sé si es válido es # para cada social centre / para paradas de metro.

Ontology:

* Tiene que cuadrar con el resource name strategy.
* Como no vamos a hacer consultas sobre líneas de autobús y nos centramos en las líneas de metro, lo quito como entidad y lo dejo como string.
  + En la explicación tendremos que decir que nos centramos en las paradas de metro.
* Sólo dejaríamos como entidades a las que atacar y hacer same\_as\_wikidata con SubwayStation, Neighborhoods y Disctricts.
* He vinculado con esta ontología de public buildings que he localizado en <https://lov.linkeddata.es/dataset/lov/terms?q=PublicBuilding> que nos los enseñó Óscar en una de las últimas clases. No sé si realmente se pueden poner ontologías.

# [Group Hands-on Assignment 3Tarea](https://moodle.upm.es/titulaciones/oficiales/mod/assign/view.php?id=47018)

The goal of this assignment is to get familiar with the cleaning and preparation of CSV data using OpenRefine.

The students will have to take their selected dataset(s) and perform the following tasks:

* **Import data.** Import data into OpenRefine.
* **Analyse and clean data.** The analysis was performed in the previous class, but it can be updated with new findings. Fix the data in order to remove errors and transform them in order to facilitate the process of generating RDF from them.

**Assignment materials**

This assignment has to be performed with the datasets selected by each group of students.

## Deliverables

The deliverables must be uploaded to the hands-on group directory in github (i.e., "/HandsOn/GroupXX/").

The deliverables required for this assignment are:

* A JSON file with the operations performed over the data for fixing and transforming them (“\*.json” under an “openrefine” directory).
* A CSV file with the updated version of the dataset (“\*-updated.csv” under a “csv” directory).
* A Markdown document with the hands-on self-assessment (”selfAssessmentHandsOn3.md" in the root of the group directory).

# [Group Hands-on Assignment 4Tarea](https://moodle.upm.es/titulaciones/oficiales/mod/assign/view.php?id=47025)

The goal of this assignment is to get familiar with the transformation of CSV data into RDF using RML mappings.

The students will have to take their selected dataset(s) and perform the following tasks:

* **Transform data into RDF**. Define RML mappings for the data and transform the data into RDF.

**Assignment materials**

* Templates for mappings are provided in the Github repository

## Deliverables

The deliverables must be uploaded to the hands-on group directory in github (i.e., "/HandsOn/GroupXX/").

The deliverables required for this assignment are:

* An RML file with the mappings (“\*.rml” under an “mappings” directory)
* A YML file with the mapping rules (optional)(“\*.yml” under an “mappings” director)
* An RDF file in the Turtle syntax with the data transformed into RDF (“\*.ttl” under an “rdf” directory).
* A SPARQL file with queries to verify your data(“queries.sparql” under an “rdf” directory).
* A Markdown document with the hands-on self-assessment (”selfAssessmentHandsOn4.md" in the root of the group directory).

# [Group Hands-on Assignment 5Tarea](https://moodle.upm.es/titulaciones/oficiales/mod/assign/view.php?id=47039)

The goal of this assignment is to get familiar with the linking of RDF data with other datasets using Open Refine.

The students will have to take their selected dataset(s) and perform the following tasks:

* **Data linking**
  + Identify those classes whose instances can be linked and identify which data sets may contain instances for the previously-identified classes.
  + Define Open Refine reconciliation services if needed.
  + Reconcile data with the identified datasets.
* **Export to RDF**
  + Update the mappings and export the data to RDF.
  + Write some SPARQL queries to check your data.

**Assignment materials**

This assignment has to be performed with the dataset(s) selected by each group of students.

## Deliverables

The deliverables must be uploaded to the hands-on group directory in github (i.e., "/HandsOn/GroupXX/").

The deliverables required for this assignment are:

* A JSON file with the operations performed over the data for linking them (“\*-with-links.json” under a “lodrefine” directory).
* A CSV file with the updated versions of the datasets (“\*-with-links.csv” under a “csv” directory).
* An RML file with the updated versions of the mappings(“\*-with-links.rml” under a “mappings” directory).
* An RDF file in the Turtle syntax with the data linked and transformed into RDF (“\*-with-links.ttl” under an “rdf” directory).
* A SPARQL file with queries to verify your links(“queries-with-links.sparql” under an “rdf” directory).

# [Group Hands-on AssignmentTarea](https://moodle.upm.es/titulaciones/oficiales/mod/assign/view.php?id=47045)

The goal of this assignment is to get familiar with the publication of linked data and ontologies.

The students will have to take their knowledge graphs and perform the following tasks:

* **Linked data publishing**
  + Publish your linked data with Helio.
  + Execute some SPARQL queries.
* **Ontology publishing (optional)**
  + Publish your ontology using Ontoology.

# [Group Hands-on Assignment 7 (optional)Tarea](https://moodle.upm.es/titulaciones/oficiales/mod/assign/view.php?id=47059)

The goal of this **OPTIONAL** assignment is to to validate the generated knowledge graph using SHACL.

The students will have to take their generated knowledge graph and perform the following tasks:

* **Define shapes using SHACL.**You can use Astrea (<https://astrea.linkeddata.es/>)**.**
* **Validate the knowledge graph according to the shapes**: use an existing validator and fix any problem found during validation.

**Assignment materials**

This assignment has to be performed with the knowledge graph generated by each group of students.

## Deliverables

The deliverables must be uploaded to the hands-on group directory in GitHub (i.e., "/HandsOn/GroupXX/").

The deliverables required for this assignment are:

* A SHACL file with the shapes ("shapes.ttl" under a "shacl" directory).
* An RDF file with the validation report (“report.ttl” under a “shacl” directory).
  + Cuando pones centros-dia-processed-without\_semicolon.csv y luego ejecutas lo del json, no da el resultado que debería. Preguntar duda. Hablar con César.

**Avancemos lo que podamos con esta parte para tenerlo claro e intentar rematarlo el viernes:**

* + Values of the ns0:belongsToMadridNeighbourhood property are not correct. Ubicar, creo que es en mapping.rml.yaml. ¿vinculado con string y número? Hay que hacer diferenciación.
  + URIs are encoded as strings. OpenRefine parece que lo identifica como URI, dónde tenemos que corregirlo.
  + ¿dónde hemos colocado el sameas para vincular con las URIs?
  + It could happen that two individuals from different classes have the same URI because the naming strategy does not ensure uniqueness. Revisar nameing strategy.
  + The property ycoord has multiple values. Creo que es en rml.ttl
  + BusStop is not a datatype.
  + Some of the properties defined with numbers as ranges are not numbers.
  + The property distrito is mixing integer and string values.
  + Boolean data values are not encoded properly.