



UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS INFORMÁTICOS

TRABAJO FIN DE MÁSTER

MÁSTER EN CIENCIA DE DATOS

CURSO 2020-2021

DATOS ABIERTOS PARA UNIVERSIDADES

AUTOR: Ángel Sánchez Fernández-Caro

TUTOR: Óscar Corcho García

Madrid, junio de 2021

Resumen

Este trabajo pretende apoyar la publicación de datos abiertos, con el estándar Núcleo Común de UniversiDATA, con una serie de conjuntos de datos diseñados para satisfacer la demanda de la sociedad actual.

A continuación, se expone el software desarrollado para la recopilación y publicación de datos para los conjuntos de datos de Núcleo Común relativos a titulaciones, centros, departamentos y presupuesto de ingresos.

Índice

Resumen	2
Introducción	1
1.1 Antecedentes	1
1.2 Sistema Integrado de Información Universitaria (SIIU)	1
Estado del arte	3
2.1 Datos abiertos	3
2.2 Legislación	3
2.3 UniversiDATA	4
2.4 Núcleo Común	4
2.4.1 Demandas	5
2.4.2 Datasets	5
2.4.3 Directrices	9
2.4.3.1 Formato de los ficheros	10
2.4.3.2 Reglas de nombrado	10
2.4.3.3 Referencias a entidades: código y descripción	10
2.5 Fuentes de información	11
2.5.1 API UPM	11
2.5.2 Registro de Universidades, Centros y Títulos (RUCT)	12
2.5.3 Página web de Universidad Politécnica de Madrid	12
2.5.4 Portal de Transparencia UPM	12
2.5.5 Instituto Nacional de Estadística (INE)	13
Objetivos	14
Metodología	15
4.1 Web Scraping	15
4.1.1 Navegador	15
4.1.2 Controlador	15
4.1 Lenguaje de programación	16
4.2 Entorno de desarrollo	16
4.3 Librerías utilizadas	16
Desarrollo	17
5.1 Titulaciones dataset	17
5.1.1 Descripción	17
5.1.2 Fuentes de información	17
5.1.3 Campos	17
5.1.4 Diagrama del software	22
5.1.5 Obtención de las variables	23
5.1.6 Consideraciones	26
5.1.7 Campos no obtenidos	26

5.1.8 Apariencia del código	27
5.2 Centros dataset	28
5.2.1 Descripción	28
5.2.2 Fuentes de información	28
5.2.3 Campos	29
5.2.4 Diagrama del software	31
5.2.5 Obtención de las variables	32
5.2.6 Consideraciones	35
5.2.7 Campos no obtenidos	35
5.2.8 Apariencia del código	35
5.3 Departamentos dataset	37
5.3.1 Descripción	37
5.3.2 Fuentes de información	37
5.3.3 Campos	37
5.3.4 Diagrama del software	39
5.3.5 Obtención de las variables	39
5.3.6 Consideraciones	40
5.3.7 Campos no obtenidos	40
5.3.8 Apariencia del código	41
5.4 Presupuesto de Ingresos dataset	42
5.4.1 Descripción	42
5.4.2 Fuentes de información	44
5.4.3 Campos	44
5.4.4 Diagrama del software	46
5.4.5 Obtención de las variables	47
5.4.6 Consideraciones	48
5.4.7 Campos no obtenidos	48
5.3.8 Apariencia del código	49
Conclusiones	50
Bibliografía	51

Índice de tablas

Tabla 1. Datasets desarrollados por Núcleo Común	6
Tabla 2. Conjunto de datasets propuestos por Núcleo Común	9
Tabla 3. Campos del dataset de Titulaciones	22
Tabla 4. Campos del dataset de Centros	31
Tabla 5. Campos del dataset de Departamentos	39
Tabla 6. Campos del dataset de Presupuesto de Ingresos	46

Índice de figuras

Figura 1. Logo UniversiDATA	4
Figura 2. Documento descargado del INE	13
Figura 3. Diagrama del software para el dataset de Titulaciones	23
Figura 4. RUCT página principal para el dataset de Titulaciones	23
Figura 5. Algunas titulaciones de la primera página	24
Figura 6. Obtención de campos para una titulación	25
Figura 7. Diagrama del software para el dataset de Centros	32
Figura 8. RUCT página principal para el dataset de Centros	32
Figura 9. Ejemplos centros de la primera página	33
Figura 10. Obtención de campos para un centro	34
Figura 11. Clarificación de la consideración para cada centro	35
Figura 12. Diagrama del software para el dataset de Departamentos	39
Figura 13. Campus Web UPM	39
Figura 14. Obtención de variables del dataset de Departamentos	40
Figura 15. Diagrama del software para el dataset de Presupuesto de Ingresos	47
Figura 16. Obtención de campos para un ingreso	47

1. Introducción

1.1 Antecedentes

Con la llegada del siglo XXI los procesos de globalización o digitalización las instituciones de educación superior han pasado a ser el centro de atención en la búsqueda de análisis y soluciones a problemas tan complejos.

Hace más de 15 años, acceder a información actualizada y contrastable del Sistema Universitario Español (SUE) era una odisea cuanto menos. La única información publicada era a través del Instituto Nacional de Estadística (INE) [1]. Esta institución recopilaba anualmente datos agregados de estudiantes y personal al servicio de las universidades y también cierta información sobre proyectos de investigación. Pero la información disponible para el público en general y para la toma de decisiones por parte de los responsables académicos no era ni mucho menos abundante.

En nuestro país, la promulgación de la Ley Orgánica de Universidades (LOU) en 2001 y su posterior modificación (LOMLOU) en 2007 supusieron un claro intento de modernización del sistema universitario español en aspectos como la apertura a la sociedad, la innovación, la empleabilidad o la adaptación de las titulaciones universitarias oficiales al Espacio Europeo de Educación Superior [2] entre otros. Una consecuencia de esta unificación es la posibilidad de evaluar las distintas titulaciones a nivel europeo con una serie de indicadores fiables.

Es en la disposición adicional decimocuarta del R.D. 1393/2007 [3] introducida por el R.D. 861/2010 [4], donde se prevé la creación y desarrollo de un sistema de información en el que participen las Universidades, las Comunidades Autónomas y el Ministerio con competencias en las Universidades. Este sistema debe dar cobertura a las necesidades de información del conjunto del sistema universitario español. También tiene como objetivo facilitar a los distintos órganos de evaluación, como la Agencia Nacional de Evaluación de la Calidad y Acreditación (ANECA) [5], la información necesaria para llevar a cabo los procedimientos relativos al seguimiento y acreditación de los títulos universitarios oficiales previstos en el real decreto.

Nos encontramos en un contexto en el que la digitalización de los órganos públicos es bastante pobre. Del mismo modo, la transparencia brilla por su ausencia y los agentes de evaluación tienen serios problemas para detectar posibles circunstancias problemáticas.

1.2 Sistema Integrado de Información Universitaria (SIIU)

A partir de entonces se empieza a desarrollar el Sistema Integrado de Información Universitaria (SIIU) en 2010 con la colaboración de las comunidades autónomas, las universidades y el ministerio. Además se crea la Comisión de Estadística e Información Universitaria, pilar fundamental del sistema. Así pues, el SIIU es una plataforma de recogida, procesamiento, análisis y difusión de datos del Sistema Universitario Español que permite disponer de información homogénea y comparable. Con este sistema se consigue avanzar en el ámbito de la transparencia para que los agentes implicados puedan tomar sus decisiones con un mayor conocimiento tras reutilizar los recursos publicados.

El SIIU se convierte en proveedor de información, no sólo de administraciones públicas, sino también de la Conferencia de Rectores de las Universidades Españolas (CRUE) [6] para la elaboración del estudio La Universidad Española en Cifras [7]. Del mismo modo, también es utilizado como fuente de datos para algunos rankings universitarios, como el elaborado por la

Fundación CyD [8], el U-Multirank [9] promovido por la Comisión Europea y el U-Ranking [10] de la Fundación BBVA y el Instituto Valenciano de Investigaciones Económicas (IVIE) [11].

El Sistema Integrado de Información Universitaria está estructurado en distintas áreas con indicadores de rendimiento para cada una de ellas que permiten hacer un estudio exhaustivo. Dichas áreas son:

- Auxiliar: estructura de las universidades, públicas y privadas, con sus centros, unidades y estudios o titulaciones impartidas [12].
- Académica: características de los estudiantes y sus variables de acceso y matrícula, así como de los procesos de EBAU y preinscripción [12].
- Recursos humanos: personal docente e investigador (PDI) y personal de administración y servicios (PAS) con sus características profesionales, sociales y demográficas [12].
- Económica: cuentas anuales de las universidades, sus liquidaciones presupuestarias y los indicadores financieros [12].
- Becas y ayudas al estudio: becas concedidas por el Estado, las comunidades autónomas y las universidades, así como información curricular y sociodemográfica de sus beneficiarios [12].
- I+D: actividad investigadora e innovadora de las universidades españolas [12].
- Inserción laboral: entrada y mantenimiento de los egresados universitarios en el mercado del trabajo [12].

Sin embargo estas herramientas no están accesibles de forma abierta. Así pues, toda la información recogida sólo podrá ser utilizada por los organismos con acceso a este sistema. De esta forma, reducimos considerablemente las posibilidades de reutilización de los datos descritos previamente.

El Ministerio de Universidades publica distintos informes y estadísticas de los ámbitos nombrados previamente [13]. Sin embargo, esta información no contiene gran cantidad de detalle y sus formatos son complejos para ser utilizados posteriormente.

2. Estado del arte

2.1 Datos abiertos

Como comentamos anteriormente, a pesar de que se publique una gran variedad de informes e infografías por parte del gobierno para temas de educación lo cierto es que la herramienta del [SIU](#) no tiene un acceso público.

El pilar fundamental de este trabajo se basa en los datos abiertos. Los datos abiertos son aquellos datos que cualquiera es libre de utilizar, reutilizar y redistribuir, con el único límite, en su caso, del requisito de atribución de su fuente o reconocimiento de su autoría [\[14\]](#).

La reutilización de la información del sector público facilita el desarrollo de nuevos productos, servicios y distintas soluciones de alto valor socioeconómico. Igualmente, permite revertir en la sociedad del conocimiento y los beneficios directos derivados de la actividad de las administraciones públicas en condiciones de transparencia. Así, la reutilización ayuda a mejorar la fiabilidad y seguridad de los datos que aquellas gestionan y contribuye a un diseño más cercano y eficiente de los servicios públicos.

2.2 Legislación

El primer paso adelante para la publicación de datos públicos de forma abierta fue lo dispuesto en el Real Decreto 861/2010 [\[4\]](#). Su finalidad es la creación de un sistema de información que dé cobertura a las necesidades del Sistema Universitario Español. Finalmente, este sistema resultó ser el Sistema Integrado de Información Universitaria (SIU), que presentamos previamente.

Con la evolución de la sociedad y las herramientas tecnológicas, la transparencia es una demanda creciente por parte de la sociedad. Con la Ley 19/2013 de Transparencia, Acceso a la Información Pública y Buen Gobierno [\[15\]](#) se dio un paso adelante para ampliar y reforzar la transparencia de la actividad pública. Del mismo modo regula y garantiza el derecho de acceso a la información y establece las obligaciones de buen gobierno que deben cumplir los responsables públicos.

A nivel europeo se ha publicado la Directiva (UE) 2019/1024 [\[16\]](#) del Parlamento Europeo y del Consejo, de 20 de junio de 2019, relativa a los datos abiertos y la reutilización de la información del sector público, sustituyendo la Directiva 2003/98/CE reformada en 2013. Mediante este texto se da un primer paso en el proceso de modernización del marco normativo de los datos abiertos y la reutilización de la información del sector público que obedecía a planteamientos propios de comienzo de siglo y, por tanto, claramente superados [\[17\]](#).

2.3 UniversiDATA

Tras la identificación de estos problemas relativos con la publicación de datos en la enseñanza pública surgió el proyecto UniversiDATA [18]. UniversiDATA es un proyecto colaborativo e impulsado por Universidades públicas españolas, a partir de una iniciativa de la empresa Dimetrical [19]. Su objetivo es impulsar los datos abiertos en el sector de la educación superior en España. Así facilitamos su uso, explotación y puesta en valor por parte tanto de los publicadores, las universidades, como a los consumidores, el sector infomediario y la sociedad en general. A partir de esta sinergia público-privada las universidades tienen la oportunidad de abrir sus datos a la sociedad y compartirlos para que puedan ser utilizados. Así pues, la participación en esta iniciativa ofrece las siguientes ventajas [20]:

- Facilita a las universidades el cumplimiento de los requerimientos planteados por la Ley de Transparencia 19/2013 [15] y por la Ley de la Comunidad de Madrid 10/2019 [21] relativos a la publicación de sus datos en formatos reutilizables.
- Provee una solución integral y estandarizada para la gestión, tratamiento, enriquecimiento, anonimización automatizada y publicación del conjunto de datos.
- Garantiza la completa disponibilidad de la solución por parte de las universidades al publicarse los datos en la plataforma de código abierto DKAN [22].
- Permite definir los contenidos de manera homogénea para todas las universidades participantes en el proyecto. La consecuencia es el valor añadido generado al abrir la posibilidad de realizar estudios comparados.
- Facilita publicar los conjuntos de datos bajo estándares internacionales como Open Fiscal Data Package [23] para los presupuestos públicos, Open Contracting Data Standard [24] para los procesos de contratación pública.
- Aumenta la visibilidad y recuperación del conjunto de datos al estar integrados tanto en el portal nacional de datos abiertos [25] como en el portal internacional de presupuestos abiertos OpenSpending [26]



Figura 1. Logo UniversiDATA

2.4 Núcleo Común

A partir de esta iniciativa UniversiDATA compuesta inicialmente por universidades y Dimetrical, se ha desarrollado un estándar de publicación de datos universitarios abiertos que se denomina el Núcleo Común. Este estándar, que comentamos ligeramente previamente, trata de dar una respuesta unificada, razonable y razonada al conjunto de datos abiertos que

cualquier universidad debe hacer públicos. Así pues, se determina los distintos datasets, campos, formatos y nivel de granularidad, por ejemplo, que deben formar estos conjuntos de datos [27]. Además, esta especificación es abierta y pública. Esto significa que cualquier universidad, independientemente de si es usuaria o no del portal UniversiDATA, y sin necesidad de pedir permiso alguno.

2.4.1 Demandas

Núcleo común ha sido diseñado atendiendo a la demanda social y legal existente de datos abiertos universitarios, y a recomendaciones específicas del sector. Para ello se han analizado diversas fuentes, entre otras [28]:

- La Ley 19/2013, de 9 de diciembre, de transparencia, acceso a la información pública y buen gobierno [15].
- El Ránking de Transparencia de Universidades de la Fundación Compromiso y Transparencia [29].
- El Ránking de Transparencia de Universidades del Dynamic Transparency Index (Dyntra) [30]. Dyntra es una plataforma colaborativa que trabaja en la medición y gestión del gobierno abierto en las organizaciones y para la sociedad civil.
- El documento 'Hacia una Universidad Abierta: Recomendaciones para el SUE' [31] de la Conferencia de Rectores de las Universidades Españolas (CRUE).

2.4.2 Datasets

En la versión inicial del [Núcleo Común](#) la comunidad [UniversiDATA](#) ha identificado 42 datasets que considera que toda universidad debería publicar para dar una cobertura mínima a la demanda legal y social de datos abiertos existente. Estos datasets están siendo definidos detalladamente en un proceso incremental y por fases, estando los primeros 13 de ellos ya disponibles como resultado de la fase 1. Los 29 restantes están en proceso de definición por parte de los participantes. Estos 13 datasets iniciales y su categoría son:

Categoría	Dataset
Estudios y Estudiantes	Titulaciones
	Acceso

	Matrículas
	Movilidad - Entrantes
	Movilidad - Salientes
Economía y Patrimonio	Presupuesto de ingresos
	Presupuestos de gastos
	Licitaciones y contratos (en curso)
Recursos Humanos	Personal (PAS)
	Personal (PDI)
	Personal (PI)
Organización y Servicios	Centros
	Departamentos

Tabla 1. Datasets desarrollados por Núcleo Común

Todos los datasets han sido clasificados en un conjunto de 6 categorías. Así pues, la clasificación de los datasets por categoría quedaría de la siguiente forma:

Categoría	Dataset
Estudios y Estudiantes	Titulaciones

	Oferta de Asignaturas
	Oferta y demanda
	Acceso
	Matrículas
	Movilidad - Entrantes
	Movilidad - Salientes
	Rendimiento académico
	Rendimiento de asignaturas
	Becas y ayudas
	Satisfacción del alumnado
	Inserción laboral
	Lectura de tesis
	Egresados
Actividad Investigadora	Grupos de investigación
	Resultados de investigación

Economía y Patrimonio	Presupuesto de ingresos
	Presupuestos de gastos
	Licitaciones y contratos (en curso)
	Cuenta de pérdidas y ganancias
	Balance de situación
	Viajes / Dietas / Gastos equipos de gobierno
	Subvenciones y ayudas públicas concedidas
	Bienes muebles / inmuebles, equipamientos e infraestructuras
	Parque móvil
	Deuda
Recursos Humanos	Personal (PAS)
	Personal (PDI)
	Personal (PI)
	Bandas salariales

	Oferta pública de empleo
	Relación de puestos de trabajo (RPT)
	Representación Sindical
	Absentismo laboral
Organización y Servicios	Centros
	Departamentos
	Otras unidades organizativas
	Agenda de los órganos de gobierno
	Convenios
	Encomiendas de gestión
	Entidades en las que la Universidad tiene participación
	Posición en rankings

Tabla 2. Conjunto de datasets propuestos por Núcleo Común [32]

2.4.3 Directrices

Cada uno de los campos de los datasets nombrados están sujetos a una serie de directrices de carácter general que se deben considerar a la hora de definir datasets que formen parte de la especificación ‘Núcleo Común’ [33].

2.4.3.1 Formato de los ficheros

El único formato obligatorio para la actual versión del Núcleo Común es Comma-Separated Values(CSV). Estos ficheros CSV seguirán las normas especificadas en el estándar RFC 4180 - Common Format and MIME Type for Comma-Separated Values (CSV) Files [34]. A resaltar especialmente los siguientes puntos [33]:

- Si bien es opcional, en Núcleo Común todos los ficheros tendrán las cabeceras de las columnas en su primera fila.
- Los nombres de las columnas no podrán incluir espacios. En aquellos casos se separarán con guiones bajos (_).
- Para facilitar su tratamiento en la mayor cantidad posible de entornos, se evitará usar vocales acentuadas y eñes en los nombres de columnas, ya que esto genera dificultades en determinadas plataformas.
- Todos los valores de todos los campos, incluyendo los nombres de las columnas, se consignarán entre comillas.

2.4.3.2 Reglas de nombrado

En cuanto al nombrado de los documentos utilizados o generados para formar los datasets, se tendrán en cuenta las siguientes directrices [33]:

- Los datasets se nombran concatenando el acrónimo de la Universidad, seguido de un guión estándar y de un nombre descriptivo del dataset. Pueden usarse espacios, mayúsculas y minúsculas. Por ejemplo, para el dataset de titulaciones quedaría de la siguiente forma:

UPM-titulaciones.csv

- Los recursos utilizados se nombran:
 - Con el mismo nombre que el dataset, todo en minúsculas, sustituyendo espacios por guiones estándar y, en su caso, seguido de un guión estándar con el periodo temporal al que hace referencia.
 - En el caso de recursos que hayan sido anonimizados llevarán el sufijo “-Anonimizado”.
 - Los recursos que estén en un formato distinto al .csv, llevarán un último sufijo compuestos por un guión estándar seguido del formato correspondiente.

2.4.3.3 Referencias a entidades: código y descripción

Todas las entidades relevantes que deben ser referidas en cualquier dataset, a excepción de las personas, deben estar representadas al menos por dos campos: un código único a nivel

de Universidad y una descripción textual. Estas entidades pueden ser centros, estudios, departamentos, etc. Se tendrá en cuenta las siguientes disposiciones [33]:

- Los campos que contengan códigos empezarán por el prefijo “cod_”.
- Los campos que contengan descripciones empezarán por el prefijo “des_”.
- Siempre que sea posible se usarán las codificaciones utilizadas en el sistema [SIIU](#) del ministerio de Educación. En su defecto nos apoyaremos en las codificaciones impuestas por el INE [1].

2.5 Fuentes de información

Para la formación de los datasets se ha utilizado información publicada de forma abierta. Se han analizado las distintas fuentes de información y finalmente se han utilizado aquellas que más se adecúan a los estándares de [Núcleo Común](#). Del mismo modo, se han descartado otras fuentes en las que se ha considerado que la información proporcionada no era del todo veraz o no estaba actualizada.

2.5.1 API UPM

En este portal se presenta información a la comunidad universitaria para que pueda ser reutilizada. Son archivos que pueden descargarse o solicitarse vía HTTP para su tratamiento. Ofrecen datos tanto públicos como privados. Sin embargo, no hemos podido obtener acceso a estos últimos.

En cuanto a datos abiertos, se ofrecen los relativos a centros, departamentos y planes de estudio de distintas titulaciones. Tras obtener y analizar esta información nos dimos cuenta que contenía algunos errores o falta de información necesaria para desarrollar el presente proyecto. Algunos de los problemas encontrados fueron:

- Falta de los códigos relativos a las entidades como pueden ser centros, departamentos o titulaciones según las directrices mostradas previamente.
- Duplicación de instancias con distintos códigos.
- Falta de gran parte de los campos necesarios para formar un dataset determinado.
- Falta de coherencia si comparamos los resultados con la página web oficial de la Universidad Politécnica de Madrid.

Por lo tanto, decidimos no continuar usando la información proporcionada por esta fuente.

2.5.2 Registro de Universidades, Centros y Títulos (RUCT)

El Registro de Universidades, Centros y Títulos [35] proporciona la información más relevante sobre las universidades, centros y títulos que conforman el sistema universitario español, en el que constan inscritos los nuevos títulos de Grado, Máster y Doctorado oficiales. Además, el RUCT tiene carácter público y de registro administrativo, y ha sido concebido como un instrumento en continua actualización. Fue creada mediante el RD 1509/2008, de 12 de septiembre, por el que se regula el Registro de Universidades, Centros y Títulos [36].

Esta herramienta nos permite obtener la mayor parte de los campos necesarios para formar los siguientes datasets:

- Centros
- Titulaciones

Para ambos datasets la información obtenida se adecúa a los estándares Núcleo Común. En caso de que se haya aplicado alguna solución distinta debido a pequeñas dudas o detalles se indica en el correspondiente apartado del dataset.

La obtención de la información se lleva a cabo utilizando la técnica de Web Scraping que más tarde explicaremos.

2.5.3 Página web de Universidad Politécnica de Madrid

Utilizando la misma técnica, Web Scraping, que para el [RUCT](#), la página web de la Universidad Politécnica de Madrid [37] se utiliza para formar el dataset relativo a los departamentos.

2.5.4 Portal de Transparencia UPM

El Portal de Transparencia [38] permite encontrar una amplia información económica y académica de la Universidad Politécnica de Madrid. Además de cumplir en todos los aspectos con la ley actual, su finalidad es proporcionar datos en abierto con una interfaz visual atractiva. Toda la información presentada se clasifica en los siguientes puntos:

- Información Institucional
- Personal
- Oferta y demanda académica
- Estudiantes
- Información Económica
- Resultados

Esta fuente se ha utilizado para formar el dataset relativo a los ingresos de la universidad que posteriormente explicaremos en detalle. Para ello se han utilizado una serie de ficheros contenidos en el apartado de 'Información Económica'.

2.5.5 Instituto Nacional de Estadística (INE)

El Instituto Nacional de Estadística [1] de España es un organismo autónomo encargado de la coordinación general de los servicios estadísticos de la Administración General del Estado (AGE) [39] y la vigilancia, control y supervisión de los procedimientos técnicos de los mismos. Entre los trabajos que realiza destacan las estadísticas sobre demografía, economía y sociedad española.

En nuestro caso, utilizaremos los ficheros contenidos en el apartado 'Relación de municipios y sus códigos por provincias. Últimos datos' [40] para obtener el código de los municipios según los estándares de Núcleo Común. El fichero descargado es el que se muestra en la imagen. Este fichero se utiliza para formar el dataset relativo a los centros.

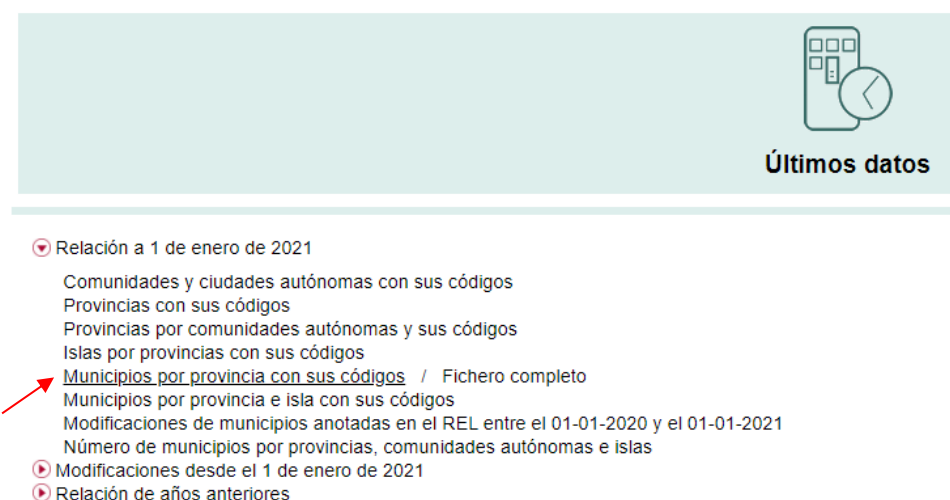


Figura 2. Documento descargado del INE

3. Objetivos

Tras presentar todo el contexto que envuelve a los datos abiertos relativos a los órganos públicos, vamos a definir los objetivos y el alcance del presente documento que está basado en el estándar Núcleo Común relativo a la publicación de datos universitarios.

Este trabajo de recopilación de datos está basado en las [directrices](#) descritas previamente de [Núcleo Común](#) junto con las especificaciones individuales de cada dataset. Cada conjunto de datos está diseñado en base a la demanda de la CRUE, así como a distintos índices de la plataforma Dyntra [\[41\]](#).

Lo que se pretende con este trabajo es desarrollar un software para que, año tras año, se publiquen los datasets relativos al nuevo curso académico. Por el momento, las publicaciones se basarán en documentos CSV. En un futuro, se puede llegar a desarrollar un portal de datos abiertos mediante la herramienta DKAN [\[22\]](#) que permita aunar en una misma plataforma la gestión de nuestro catálogo de datos y la gestión de nuestro contenido online.

4. Metodología

Una vez presentados los objetivos junto con las distintas fuentes de datos, vamos a presentar de manera general qué herramientas hemos utilizado para desarrollar el presente trabajo.

4.1 Web Scraping

Para gran parte de la generación de los datasets hemos utilizado una técnica llamada web scraping. Se trata de un proceso de usar bots para extraer contenido y datos, el código HTML, de un sitio web. Es decir, el web scraping consiste en navegar automáticamente por una web y extraer información de ella. Al software programado para hacer esta función es a lo que se le llama bot. Finalmente, los datos obtenidos pueden ser guardados en bases de datos o en formatos como CSV, XLS, etc.

El scraping es legal siempre y cuando los datos recabados estén disponibles libremente para terceros en la web. Sin embargo, esta técnica no siempre es legal. Se debe tener en cuenta los derechos de propiedad intelectual de los sitios web con el fin de evitar cualquier problema legal.

Para realizar el scraping o raspado de datos vamos a utilizar la librería Selenium [\[42\]](#). La versión actual de Selenium está basada exclusivamente en HTML y JavaScript y permite a los desarrolladores probar y registrar las interacciones con una aplicación web y luego repetirlas las veces que se desee, de forma completamente automática. Además, esta herramienta es de código abierto, con lo que no hay costes de licencia. Para simular dichas acciones del usuario con cualquier navegador utilizaremos la interfaz Selenium WebDriver. Desde el 2018, la API es un estándar W3C oficial [\[43\]](#).

4.1.1 Navegador

Selenium puede ser utilizado mediante distintos navegadores. En nuestro caso, utilizaremos Google Chrome [\[44\]](#) con la versión 91.0.4472.124 de 64 bits.

4.1.2 Controlador

El siguiente paso es descargar e instalar el controlador del navegador al que se vaya a conectar Selenium, en nuestro caso para Google Chrome. El controlador necesario para poder automatizar este navegador es ChromeDriver, que podemos descargar desde su página oficial [\[45\]](#) la versión necesaria para nuestro sistema. En nuestro caso y teniendo en cuenta la versión de nuestro navegador, nos descargamos el ChromeDriver 91.0.4472.101.

4.1 Lenguaje de programación

Este proyecto ha sido desarrollado empleando el lenguaje de programación Python [\[46\]](#) en un sistema operativo de Windows 10. Concretamente, la versión utilizada de Python es python 3.9.1 de 64 bit [\[47\]](#).

4.2 Entorno de desarrollo

El entorno de desarrollo utilizado es Sublime Text 3 [\[48\]](#) para Windows 10 de 64 bits. Sublime Text es un editor de texto que nos permite ejecutar códigos en una gran variedad de lenguajes de programación.

4.3 Librerías utilizadas

Para poder realizar la obtención de datos y su tratamiento se han utilizado distintas librerías. En concreto, estas librerías son:

- Selenium Webdriver [\[49\]](#): utilizado para realizar el web scraping. Licencia Apache 2.0 [\[50\]](#).
- Pandas [\[51\]](#): utilizada para la manipulación y tratamiento de datos. Licencia BSD-3-Clause [\[52\]](#).
- Regular expression - re [\[53\]](#): nos permite la búsqueda de una subcadena dentro de un texto. Licencia CNRI-Python [\[54\]](#).
- Unicodedata [\[55\]](#): proporciona acceso a la base de datos de caracteres Unicode (UCD), que define las propiedades de todos los caracteres Unicode. Nos permite solucionar problemas relativos a los acentos o las eñes. Licencia Python-2.0 [\[56\]](#).
- Geopy [\[57\]](#): facilita localizar las coordenadas de direcciones, ciudades, países y puntos de referencia en todo el mundo mediante geocodificadores de terceros y otras fuentes de datos. Licencia MIT [\[58\]](#).

5. Desarrollo

En el presente punto vamos a explicar en detalle todo lo relativo a la formación de los datasets.

5.1 Titulaciones dataset

5.1.1 Descripción

Este dataset contiene todas las titulaciones ofertadas por la universidad, en nuestro caso por la Universidad Politécnica de Madrid. Se incluirán todas las titulaciones de grado, máster y doctorado. Sin embargo, otras como pueden ser títulos propios o cursos cortos no están incluidas [59].

Este [dataset](#) está incluido en la categoría ‘Estudios y estudiantes’. Su frecuencia de actualización es anual, siendo el mes de Marzo cuando se realiza dicha actualización. Este dataset no aplica ninguna técnica de anonimizado puesto que para su formación no incluye ningún campo que pueda ser un dato personal y por lo tanto, que esté referido a una persona. Para su diseño se han tenido en cuenta:

- Recomendaciones CRUE [31].
- Indicador due171 de Dyntra [60]. Este indicador hace referencia a la publicación de la oferta de títulos de grado y de postgrado.

5.1.2 Fuentes de información

En este dataset sólo se ha utilizado como fuente de información la herramienta RUCT.

5.1.3 Campos

Este dataset está compuesto por un total de 34 variables que suelen ir emparejadas por un código y su descripción.

Campo	Descripción	Observaciones
curso_academico	curso académico al que se refieren los datos	
cod_universidad	código de la universidad	

des_universidad	nombre de la universidad	
cod_tipo_estudio	código del tipo de estudio	<ul style="list-style-type: none"> • 'G': para grados • 'M': para másteres • 'D': para doctorados
des_tipo_estudio	nombre del tipo de estudio	<ul style="list-style-type: none"> • 'Grado' • 'Master' • 'Doctorado'
cod_tipo_unidad_responsable	código del tipo de unidad responsable del estudio	<ul style="list-style-type: none"> • '1' para Centro • '2' para Departamento • '3' para Otra unidad
des_tipo_unidad_responsable	nombre del tipo de unidad responsable	<ul style="list-style-type: none"> • 'Centro' • 'Departamento' • 'Otra'
cod_unidad_responsable	código de la unidad responsable	
des_unidad_responsable	nombre de la unidad responsable	
cod_titulacion	código de la titulación	
des_titulacion	nombre de la titulación	
cod_rama	código de la rama de enseñanza	<ul style="list-style-type: none"> • '1' para Artes y Humanidades • '2' para Ciencias • '3' para Ciencias Sociales y Jurídicas • '4' para Ingeniería y Arquitectura • '5' para Ciencias de la Salud
des_rama	nombre de la rama de enseñanza	<ul style="list-style-type: none"> • 'Artes y Humanidades' • 'Ciencias'

		<ul style="list-style-type: none"> • 'Ciencias Sociales y Jurídicas' • 'Ingeniería y Arquitectura' • 'Ciencias de la Salud'
cod_interuniversitario	¿Se trata de un estudio interuniversitario?	<ul style="list-style-type: none"> • '0' para No es interuniversitario • '1' para Sí es interuniversitario con alguna Universidad española • '2' para Sí es interuniversitario, con alguna Universidad extranjera (aunque participen universidades españolas)
des_interuniversitario	descripción de si es interuniversitario	<ul style="list-style-type: none"> • ' No es interuniversitario' • 'Sí es interuniversitario con alguna Universidad española' • 'Sí es interuniversitario, con alguna Universidad extranjera'
cod_rol_coordinacion	si es interuniversitario, ¿es la universidad la coordinadora?	<ul style="list-style-type: none"> • '0' para No es coordinadora • '1' para Sí es coordinadora • '8' para No aplica coordinación
des_rol_coordinacion	descripción del rol de coordinación	<ul style="list-style-type: none"> • 'No es coordinadora' • 'Si es coordinadora' • 'No aplica coordinación'

cod_impartido	código de si es impartido en el curso de referencia	<ul style="list-style-type: none"> • '0' si No se imparte en el curso académico de referencia • '1' si Se imparte en el curso académico de referencia
des_impartido	descripción de si es impartido en el curso de referencia	<ul style="list-style-type: none"> • 'No se imparte en el curso académico de referencia' • 'Se imparte en el curso académico de referencia'
anio_inicio	año de inicio de la impartición del curso de referencia	
cod_situacion_actual	código de la situación de la titulación en el curso de referencia	<ul style="list-style-type: none"> • '1' para Activo (la universidad oferta plazas) • '2' para Inactivo (No está activo este curso) • '3' para En proceso de extinción (No se oferta para estudiantes de nuevo ingreso y sí para el resto de estudiantes) • '4' para Extinguido este curso (En algún momento ha estado activo y ya no tiene matrícula)
des_situacion_actual	descripción de la situación de la titulación en el curso de referencia	<ul style="list-style-type: none"> • 'Activo' • 'Inactivo' • 'En proceso de extinción' • 'Extinguido este curso'

cod_grado_presencialidad	grado de presencialidad del estudio	<ul style="list-style-type: none"> • '1' para Estudio presencial • '2' para Estudio semipresencial • '3' para Estudio no presencial • '4' para Varias modalidades <p>* Si no aplica, dejar vacío.</p>
des_grado_presencialidad	descripción del grado de presencialidad del estudio	<ul style="list-style-type: none"> • 'Estudio presencial' • 'Estudio semipresencial' • 'Estudio no presencial' • 'Varias modalidades'
cod_idioma_extranjero	código de si la titulación se oferta en idioma extranjero	<ul style="list-style-type: none"> • '0' para No • '1' para Inglés • '2' para Sí, existen planes separados, uno de ellos en inglés • '3' para Sí, otro idioma extranjero • '4' para Sí, existen planes separados, uno de ellos en otro idioma extranjero
des_idioma_extranjero	descripción de si la titulación se oferta en idioma extranjero	<ul style="list-style-type: none"> • 'No' • 'Inglés' • 'Sí, existen planes separados, uno de ellos en inglés' • 'Sí, existen planes separados, uno de ellos en otro idioma extranjero'
precio_credito_1	precio del crédito en 1ª matrícula	

precio_credito_2	precio del crédito en 2ª matrícula	
precio_credito_3	precio del crédito en 3ª matrícula	
precio_credito_4	precio del crédito en 4ª matrícula	
precio_credito_5	precio del crédito en 5ª matrícula	
precio_tutela	en caso de doctorados, precio del curso	
num_creditos_necesarios	número total de créditos que el estudiante debe superar para obtener el título	
num_creditos_ofertados	número total de créditos que oferta la universidad en esta titulación en el curso académico de referencia	

Tabla 3. Campos del dataset de Titulaciones [59]

5.1.4 Diagrama del software

El diagrama para la obtención de este conjunto de datos es bastante sencillo. Partimos de la plataforma [RUCT](#). Seleccionamos las pestañas adecuadas para obtener los datos relativos a grados, máster y doctorados.

Mediante web scraping obtenemos todos los campos posibles para cada una de las titulaciones mostradas en la herramienta.

A continuación, adecuamos los campos almacenados a las [directrices](#) generales de Núcleo Común y particulares de este dataset.

Finalmente, guardamos nuestro conjunto de datos en un fichero de formato CSV.



Figura 3. Diagrama del software para el dataset de Titulaciones

5.1.5 Obtención de las variables

Para la obtención de las variables hemos utilizado [web scraping](#). A continuación, mostraremos la forma mediante la que hemos obtenido las variables.

Inicialmente, la aplicación selecciona la universidad de la que se desea obtener los datos. Del mismo modo selecciona el tipo de estudio, ya sea grado, máster o doctorado. Esta etapa inicial finaliza pulsando el botón 'Consultar'.

Figura 4. RUCT página principal para el dataset de Titulaciones

A continuación, nos aparecen todas las titulaciones que coinciden con los valores seleccionados previamente. Las titulaciones están alojadas en una tabla HTML y se distribuyen en una serie de páginas con 25 titulaciones cada una.

De esta primera tabla se obtienen un total de 8 campos. El resto de campos son obtenidos tras pulsar en los 2 hipervínculos que tiene cada titulación, si lo tiene. El primero ha sido descrito como 'Descripción' y el segundo como 'Detalle' en la Figura 6








Código	Título	Universidad	Nivel académico	Estado	
2503225	Graduado o Graduada en Administración y Dirección de Empresas por la Universidad Politécnica de Madrid	Universidad Politécnica de Madrid	Grado	Publicado en B.O.E.	
2501481	Graduado o Graduada en Arquitectura Naval por la Universidad Politécnica de Madrid	Universidad Politécnica de Madrid	Grado	Publicado en B.O.E. (TITULACIÓN RENOVADA)	
2502648	Graduado o Graduada en Biotecnología por la Universidad Politécnica de Madrid	Universidad Politécnica de Madrid	Grado	Publicado en B.O.E. (TITULACIÓN RENOVADA)	
2503943	Graduado o Graduada en Ciencia de Datos e Inteligencia Artificial por la Universidad Politécnica de Madrid	Universidad Politécnica de Madrid	Grado	Publicado en B.O.E.	
2503468	Graduado o Graduada en Ciencias Agrarias y Bioeconomía por la Universidad Politécnica de Madrid	Universidad Politécnica de Madrid	Grado	Publicado en B.O.E.	
2500734	Graduado o Graduada en Ciencias de la Actividad Física y del Deporte por la Universidad Politécnica de Madrid	Universidad Politécnica de Madrid	Grado	Publicado en B.O.E. (TITULACIÓN RENOVADA)	
2503020	Graduado o Graduada en Diseño de Moda por la Universidad Politécnica de Madrid	Universidad Politécnica de Madrid	Grado	Publicado en B.O.E.	
2500800	Graduado o Graduada en Edificación por la Universidad Politécnica de Madrid	Universidad Politécnica de Madrid	Grado	Publicado en B.O.E. (TITULACIÓN RENOVADA POR ACREDITACIÓN INSTITUCIONAL)	
2502137	Graduado o Graduada en Fundamentos de la Arquitectura por la Universidad Politécnica de Madrid	Universidad Politécnica de Madrid	Grado	Publicado en B.O.E. (TITULACIÓN RENOVADA)	
2503799	Graduado o Graduada en Gestión y Operaciones del Transporte Aéreo por la Universidad Politécnica de Madrid	Universidad Politécnica de Madrid	Grado	Publicado en B.O.E. (TITULACIÓN RENOVADA POR ACREDITACIÓN INSTITUCIONAL)	

Figura 5. Algunas titulaciones de la primera página

En el siguiente esquema se puede apreciar de una forma simplificada el funcionamiento de la aplicación de web scraping. Se puede comprobar cómo la aplicación navega automáticamente para cada una de las titulaciones durante todas las páginas para cada tipo de titulación.

En la parte superior derecha están los campos que permanecen constantes para cada uno de los centros.

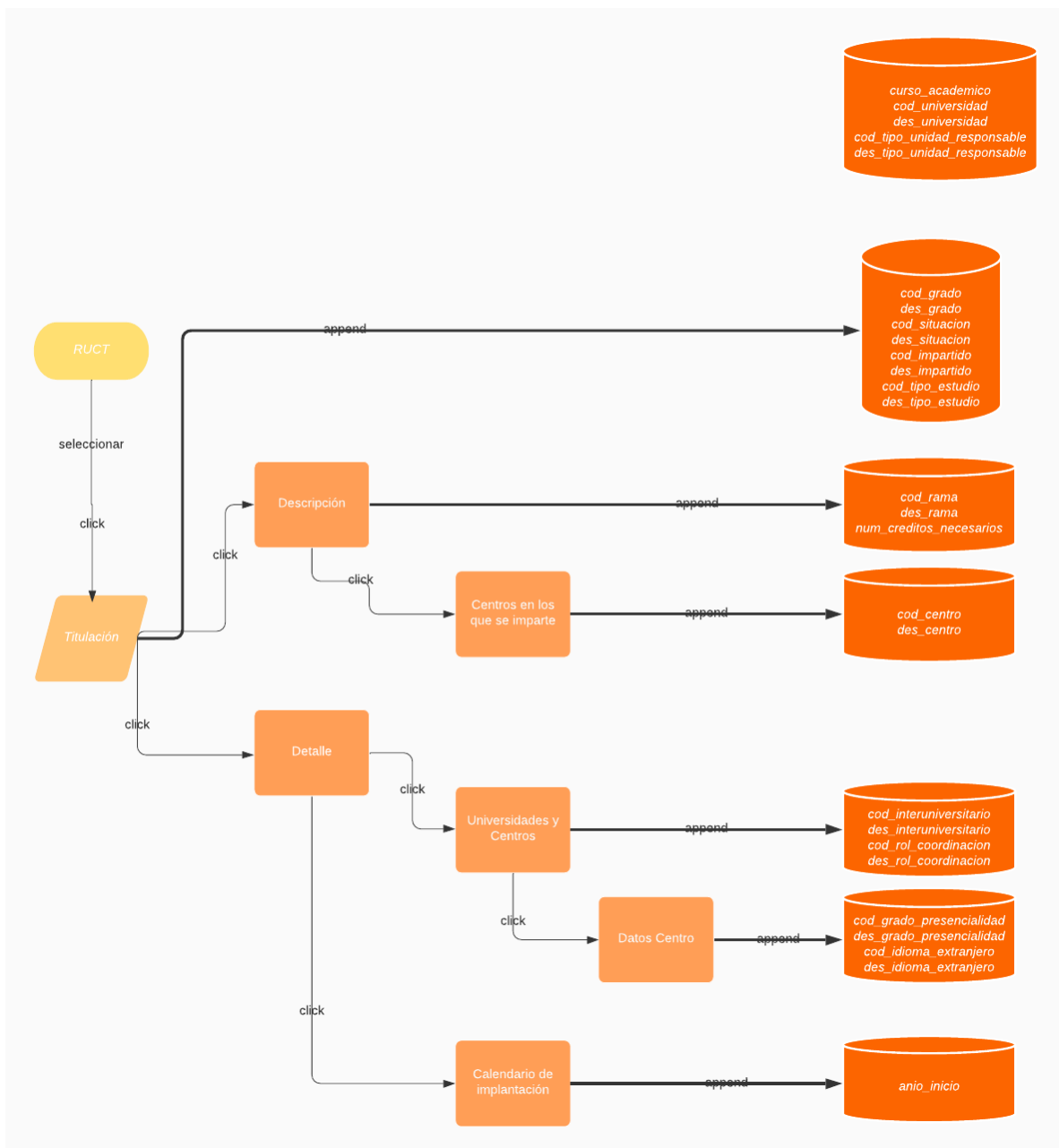


Figura 6. Obtención de campos para una titulación

Para cada titulación se realizan una serie de procesos:

- Inicialmente se obtienen una serie de campos directamente o deducidos de la tabla inicial presentada para cada titulación.
- A continuación, se pulsa el hipervínculo de la titulación
 - De la pestaña 'Descripción' se obtienen una serie de campos
 - De la pestaña 'Centros en los que se imparte' se obtienen otros campos.
- Por último se pulsa en el hipervínculo de la última columna, si lo hay.

- Pulsamos en el menú 'Universidades y Centros' y guardamos una serie de valores.
 - Pulsamos en 'Datos Centro' (símbolo de lupa) y obtenemos otros campos.
- Pulsamos en el menú 'Calendario de implantación' y obtenemos el último valor.

5.1.6 Consideraciones

Para algunos de los campos que conforman el dataset se han tenido en cuenta una serie de consideraciones para solucionar ciertos problemas. Estas consideraciones son las siguientes:

- **cod_tipo_unidad_responsable:** en este campo al únicamente poder obtener el centro de la herramienta RUCT, hemos considerado que para todas las titulaciones la unidad responsable será: '1'. Se aplicará el valor correspondiente para el campo **des_tipo_unidad_responsable**.
- **cod_impartido:** el criterio establecido para definir a una titulación como impartida es si tiene alumnos matriculados en el curso académico de referencia. Puesto que no tenemos acceso a las matrículas de los alumnos, se ha considerado que la titulación no se imparte cuando el estado del estudio es: 'Titulación Extinguida'. Se aplicará el valor correspondiente para el campo **des_impartido**.
- **cod_situacion_actual:** no se ha tenido en cuenta el código '2' puesto que no tenemos acceso a datos históricos para saber si una titulación está inactiva temporalmente.
- **cod_idioma_extranjero:** el criterio para definir si una titulación es 'idioma extranjero' es que al menos el 50% de los créditos que se ofertan son en uno o varios idiomas extranjeros. Puesto que no tenemos acceso al plan de estudio de cada titulación, el idioma se obtendrá del especificado en la herramienta RUCT para cada titulación.

Por otro lado, se ha tenido en cuenta las variaciones que aparecen en el HTML de cada titulación a la hora de acceder al elemento deseado. Con esto conseguimos una mayor veracidad a la hora de almacenar el valor final. Del mismo modo, se ha tratado con distintos errores a la hora de acceder a un elemento. Así conseguimos agregar un valor nulo y la aplicación sigue su curso de ejecución automática.

5.1.7 Campos no obtenidos

Sin embargo, no todos los [campos](#) han podido ser obtenidos utilizando datos abiertos. En caso de que un campo no haya podido ser obtenido, su valor se reemplazará por un espacio en blanco ' '. Los campos no obtenidos son los siguientes:

- **precio_credito_1**

- precio_credito_2
- precio_credito_3
- precio_credito_4
- precio_credito_5
- precio_tutela
- num_creditos_ofertados

5.1.8 Apariencia del código

En lo que respecta al código elaborado la estructura es la siguiente. Primeramente, configuramos el controlador y accedemos a la página web. A continuación, en cada menú desplegable seleccionamos la opción deseada mediante el método 'find_elements_by_xpath' y finalmente pulsamos con el método 'click()'.

```
driver_path = 'C:\\Users\\sanch\\Downloads\\chromedriver.exe'
option = Options()
option.add_argument('--headless') #ocultar pestaña navegador
#TITULACIONES - GRADO
url = 'https://www.educacion.gob.es/ruct/consultaestudios?actual=estudios'
driver = webdriver.Chrome(driver_path, options=option)
driver.get(url)

uni = driver.find_elements_by_xpath('//*[ @id="codigoUniversidad" ]')
uni[0].click()
UPM = driver.find_elements_by_xpath('//*[ @id="codigoUniversidad" ]/option[79]')
UPM[0].click()

nivel_academico = driver.find_elements_by_xpath('//*[ @id="codigoTipo" ]')
nivel_academico[0].click()
grado = driver.find_elements_by_xpath('//*[ @id="codigoTipo" ]/option[4]')
grado[0].click()
buscar =
driver.find_elements_by_xpath('//*[ @id="consultaestudios_listaestudios" ]')
buscar[0].click()
```

A continuación, obtenemos algunos campos de la página inicial de titulaciones como se muestra en la Figura 5.


```
grados = driver.find_elements_by_xpath('//*[@id="estudio"]/tbody/tr')
links =
len(driver.find_elements_by_xpath('//*[@id="estudio"]/tbody/tr/td[2]/a'))

for grado in grados:
    cod_grado.append(grado.find_elements_by_xpath('..//td[1]')[0].text)
    nombre_grado.append(grado.find_elements_by_xpath('..//td[2]')[0].text)
    estado_grado.append(grado.find_elements_by_xpath('..//td[5]')[0].text)
    cod_tipo_estudio.append('G')
    des_tipo_estudio.append('Grado')
```

Posteriormente con la misma metodología accedemos a los enlaces como muestra la Figura 6 y obtenemos el resto de campos de una forma más compleja y tratando distintos errores.

5.2 Centros dataset

5.2.1 Descripción

Este dataset contiene todos los centros de la universidad, en nuestro caso de la Universidad Politécnica de Madrid, que imparten docencia o tienen personal investigador asociado. Estos centros pueden ser escuelas, facultades, centros de estudios, institutos de investigación, hospitales, etc. No se incluirán los departamentos, ya que para ellos ya hay un dataset específico definido [61].

Este dataset está incluido en la categoría de ‘Organización y Servicios’. Su frecuencia de actualización es anual, siendo el mes de Marzo cuando se realiza dicha actualización. Este conjunto de datos no aplica ninguna técnica de anonimizado puesto que para su formación no incluye ningún campo que pueda ser un dato personal y por lo tanto, que esté referido a una persona. Para su diseño se han tenido en cuenta:

- Recomendaciones CRUE [31].

5.2.2 Fuentes de información

En este dataset se han utilizado varias fuentes de información. Las fuentes en cuestión son las siguientes:

- Mayoritariamente los datos se han obtenido utilizando la herramienta [RUCT](#) mediante [web scraping](#).
- Mediante el [INE](#) hemos descargado un fichero para obtener el código de los municipios en la Comunidad de Madrid.

- Mediante la librería [Geopy](#) hemos obtenido los campos de la latitud y longitud de los municipios en los que se sitúa el centro.

5.2.3 Campos

Este dataset está compuesto por un total de 18 variables que suelen ir emparejadas por un código y su descripción.

Campos	Descripción	Observaciones
curso_academico	curso académico al cual se refieren los datos	
cod_universidad	código de la universidad	
des_universidad	nombre de la universidad	
cod_centro	código del centro	
des_centro	nombre del centro	
cod_municipio	código del municipio en el que está ubicado dicho centro	* En este código sólo se incluye el código del municipio. Por tanto, no está incluido el código de la comunidad autónoma en la que está situado dicho municipio.
des_municipio	nombre del municipio en el que está ubicado dicho centro	
cod_campus	código del campus al que está asociado dicho centro	
des_campus	nombre del campus al que está asociado dicho centro	

cod_tipo_centro	código del tipo de centro	<ul style="list-style-type: none"> • 'C1' para Centro propio • 'C2' para Centro adscrito • 'C3' para Centro asociado / vinculado • 'O1' para Instituto universitario de investigación • 'O2' para Escuela de doctorado • 'O3' para Hospital • 'O4' para Fundación • 'XX' para Otros
des_tipo_centro	nombre del tipo de centro	<ul style="list-style-type: none"> • 'Centro propio' • 'Centro adscrito' • 'Centro asociado / vinculado' • 'Instituto universitario de investigación' • 'Escuela de doctorado' • 'Hospital' • 'Fundación' • 'Otros'
cod_naturaleza_centro	código de la naturaleza del centro	<ul style="list-style-type: none"> • '1' para Público • '2' para Privado • '3' para Privado de la Iglesia • '4' para Mixto
des_naturaleza_centro	nombre de la naturaleza del centro	<ul style="list-style-type: none"> • 'Público' • 'Privado' • 'Privado de la Iglesia' • 'Mixto'
cod_situacion_centro	código de la situación del centro a la fecha de referencia del dataset	<ul style="list-style-type: none"> • '1' para Activo • '2' para Inactivo • '3' para En proceso de extinción • '4' para Extinguido en el curso de referencia

des_situacion_centro	nombre de la situación del centro a la fecha de referencia del dataset	<ul style="list-style-type: none"> • 'Activo' • 'Inactivo' • 'En proceso de extinción' • 'Entinguido en el curso de referencia'
fecha_desde_situacion_centro	fecha desde la cual el centro se encuentra en la situación consignada	
latitud	latitud de la geolocalización del centro	
longitud	longitud de la geolocalización del centro	

Tabla 4. Campos del dataset de Centros [61]

5.2.4 Diagrama del software

El diagrama para la obtención de este conjunto de datos es bastante sencillo. Partimos de la plataforma [RUCT](#). Seleccionamos las pestañas adecuadas para obtener los datos relativos a los centros.

Mediante web scraping obtenemos todos los campos posibles para cada una de los centros mostrados en la herramienta. Utilizamos la librería 'geopy' para recopilar las latitudes y longitudes de los municipios en los que se sitúan los centros.

A continuación, adecuamos los campos almacenados a las [directrices](#) generales de Núcleo Común y particulares de este dataset. Integramos el fichero descargado del [INE](#) [40] para obtener el campo relativo al código del municipio.

Finalmente, guardamos nuestro conjunto de datos en un fichero de formato CSV.

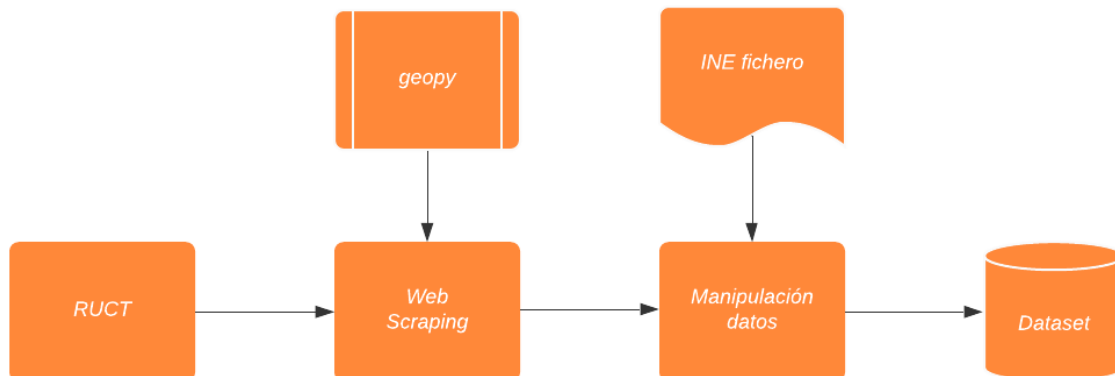


Figura 7. Diagrama del software para el dataset de Centros

5.2.5 Obtención de las variables

Para la obtención de las variables hemos utilizado web scraping mayoritariamente. A continuación, mostraremos la forma mediante la que hemos obtenido las variables.

Inicialmente, la aplicación selecciona la universidad de la que se desea obtener los datos. Esta etapa inicial finaliza pulsando el botón 'Consultar'.

Registro de Universidades, Centros y Títulos (RUCT)

Sección universidades
Sección centros
Sección títulos

Consulta de centros Contacte

* Para elegir un centro, debe seleccionar una

Código del centro:

Denominación del centro:

Universidad:

Centro:

Naturaleza vinculación:

Comunidad Autónoma:

Figura 8. RUCT página principal para el dataset de Centros

A continuación, nos aparecen todos los centros que coinciden con los valores seleccionados previamente. Los centros están alojados en una tabla HTML y se distribuyen en una serie de páginas con hasta 25 centros en cada una.

De esta primera tabla se obtienen un total de 2 campos. El resto de campos son obtenidos tras pulsar en el hipervínculo de la columna 'Universidad' que tiene cada centro. Este enlace ha sido descrito como 'Descripción' en la figura 10.

Código Univ.	Universidad	Código Centro	Centro	Títulos	Histórico	Acreditación Institucional
025	Universidad Politécnica de Madrid	28053654	Centro de Investigación en Biotecnología y Genómica de Plantas (CBGP)	 Ver títulos	 Ver histórico	
025	Universidad Politécnica de Madrid	28053836	Centro Superior de Diseño de Moda de Madrid	 Ver títulos	 Ver histórico	
025	Universidad Politécnica de Madrid	28053940	Escuela Internacional de Doctorado de la Universidad Politécnica de Madrid	 Ver títulos	 Ver histórico	
025	Universidad Politécnica de Madrid	28048300	Escuela Politécnica de Enseñanza Superior	 Ver títulos	 Ver histórico	
025	Universidad Politécnica de Madrid	28026870	Escuela Técnica Superior de Arquitectura	 Ver títulos	 Ver histórico	
025	Universidad Politécnica de Madrid	28026729	Escuela Técnica Superior de Edificación	 Ver títulos	 Ver histórico	Acreditación Institucional(19/07/2019)
025	Universidad Politécnica de Madrid	28051761	Escuela Técnica Superior de Ingeniería Aeronáutica y del Espacio	 Ver títulos	 Ver histórico	Acreditación Institucional(11/03/2021)
025	Universidad Politécnica de Madrid	28053885	Escuela Técnica Superior de Ingeniería Agronómica, Alimentaria y de Biosistemas	 Ver títulos	 Ver histórico	

Figura 9. Ejemplos centros de la primera página

En el siguiente esquema se puede apreciar de una forma simplificada el funcionamiento de la aplicación de web scraping. Se puede comprobar cómo la aplicación navega automáticamente para cada uno de los centros durante todas las páginas.

En la parte superior derecha están los campos que permanecen constantes para cada uno de los centros.

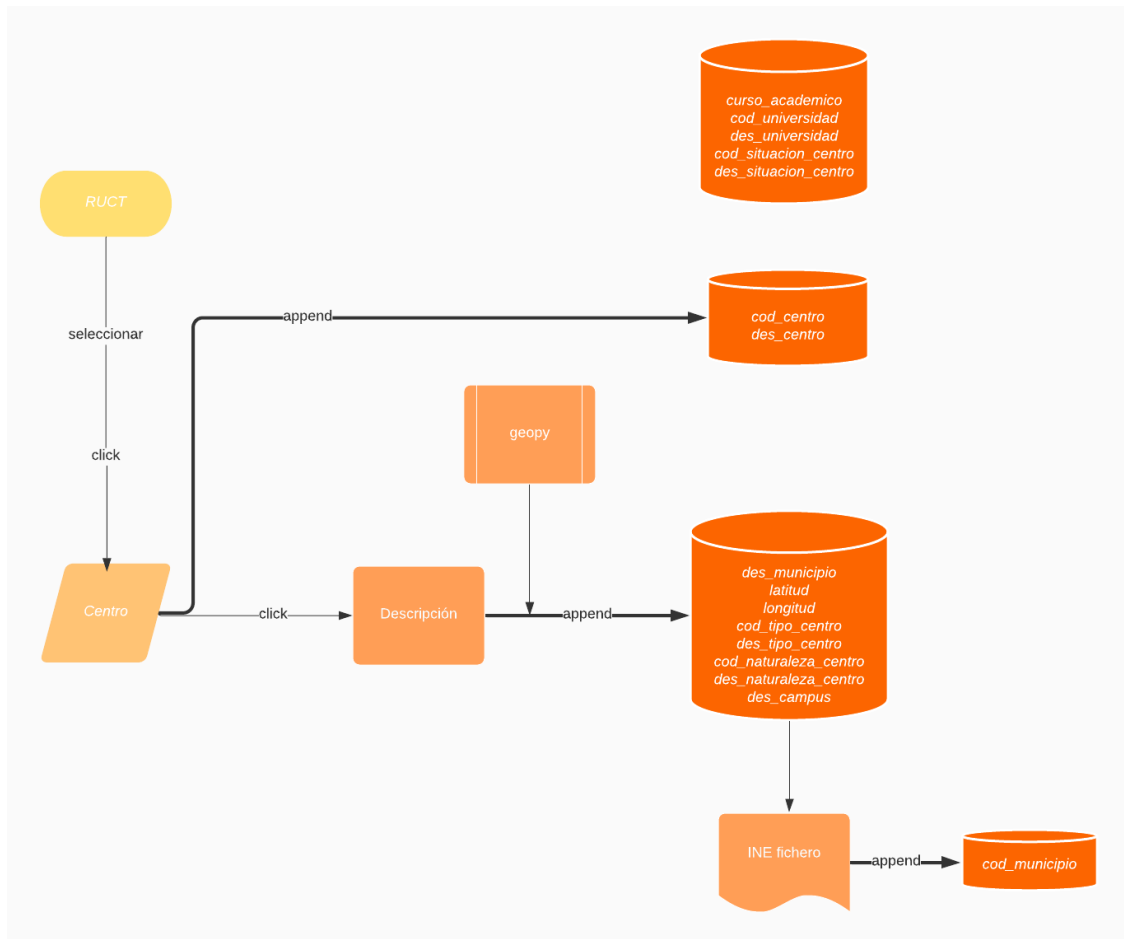


Figura 10. Obtención de campos para un centro

Para cada centro se realizan una serie de procesos:

- Inicialmente se obtienen una serie de campos directamente o deducidos de la tabla inicial presentada para cada centro.
- A continuación, se pulsa el hipervínculo del centro
 - De la pestaña 'Descripción' se obtienen una serie de campos
 - Con la ayuda de la librería se obtiene la localización geográfica de los municipios donde se sitúa cada centro.
 - Utilizando este municipio junto con el archivo descargado del [INE](#), obtenemos el código para cada municipio.

5.2.6 Consideraciones

Para algunos de los [campos](#) que conforman el dataset se han tenido en cuenta una serie de consideraciones para solucionar ciertos problemas. Estas consideraciones son las siguientes:

- **cod_tipo_centro:** para cada uno de los centros se pueden obtener varias opciones para este campo. En nuestro caso, hemos obtenido los valores que corresponden al campo 'Calificación jurídica' que se puede apreciar en la imagen. Se ha descartado la opción relativa al campo 'Tipo de centro' de la imagen por determinadas situaciones en las que el valor no se encontraba en las opciones propuestas por Núcleo Común en el diseño del dataset. Se aplicará el valor correspondiente para el campo **des_tipo_centro**.

Datos de identificación

Código del centro : 28026778
Tipo de centro : Escuela Técnica Superior
Calificación jurídica : Público
Naturaleza vinculación : Propio

Figura 11. Clarificación de la consideración para cada centro

- **cod_situacion_centro:** se ha supuesto que todos los centros están activos, puesto que no aparece información que indique lo contrario como pasaba en el caso de las titulaciones. Se aplicará el valor correspondiente para el campo **des_situacion_centro**.
- **latitud:** la latitud obtenida ha sido la del municipio del centro, en lugar de la del propio centro en sí.
- **longitud:** la longitud obtenida ha sido la del municipio del centro, en lugar de la del propio centro en sí.

5.2.7 Campos no obtenidos

Sin embargo, no todos los [campos](#) han podido ser obtenidos utilizando datos abiertos. En caso de que un campo no haya podido ser obtenido, su valor se reemplazará por un espacio en blanco ' '. Los campos no obtenidos son los siguientes:

- **cod_campus**
- **fecha_desde_situacion_centro**

5.2.8 Apariencia del código

En lo que respecta al código elaborado la estructura es la siguiente.

Primeramente, configuramos el controlador y accedemos a la página web. A continuación, en cada menú desplegable seleccionamos la opción deseada mediante el método 'find_elements_by_xpath' y finalmente pulsamos con el método 'click()'.

```
driver_path = 'C:\\Users\\sanch\\Downloads\\chromedriver.exe'
option = Options()
option.add_argument('--headless') #ocultar pestaña navegador

#CENTROS
url =
'https://www.educacion.gob.es/ruct/consultacentros.action?actual=centros'
driver = webdriver.Chrome(driver_path, options=option)
driver.get(url)

uni = driver.find_elements_by_xpath('//*[@id="codigoUniversidad"]')
uni[0].click()
UPM =
driver.find_elements_by_xpath('//*[@id="codigoUniversidad"]/option[79]')
UPM[0].click()
buscar =
driver.find_elements_by_xpath('//*[@id="consultacentros_listacentros"]')
buscar[0].click()
```

A continuación, obtenemos algunos campos de la página inicial de titulaciones como se muestra en la Figura 9.

```
centros = driver.find_elements_by_xpath('//*[@id="centro"]/tbody/tr')

for centro in centros:
    cod_centro.append(centro.find_ele-
ments_by_xpath('..//td[3]')[0].text)
    nombre_centro.append(centro.find_ele-
ments_by_xpath('..//td[4]/a')[0].text)
```

Posteriormente con la misma metodología accedemos a los enlaces como muestra la Figura 10 y obtenemos el resto de campos de una forma muy similar sin problemas.

5.3 Departamentos dataset

5.3.1 Descripción

Este dataset contiene todos los departamentos de la Universidad Politécnica de Madrid. En este caso se entiende por departamento cualquier unidad organizativa docente o investigadora, no como unidad administrativa general [62].

Este dataset está incluido en la categoría de 'Organización y servicios'. Su frecuencia de actualización es anual, siendo el mes de Marzo cuando se realiza dicha actualización. Este conjunto de datos no aplica ninguna técnica de anonimizado puesto que para su formación no incluye ningún campo que pueda ser un dato personal y por lo tanto, que esté referido a una persona. Para su diseño se han seguido:

- Recomendaciones CRUE [31].

5.3.2 Fuentes de información

En este dataset sólo se ha utilizado como fuente de información la [página web](#) de la Universidad Politécnica de Madrid.

5.3.3 Campos

Este dataset está compuesto por un total de 12 variables que suelen ir emparejadas por un código y su descripción.

Campos	Descripción	Observaciones
curso_academico	curso académico al cual se refieren los datos	
cod_universidad	código de la universidad	
des_universidad	nombre de la universidad	
cod_departamento	código del departamento	

des_departamento	nombre del departamento	
cod_municipio	código del municipio en el que está ubicado dicho departamento	
des_municipio	nombre del municipio en el que está ubicado dicho departamento	
cod_campus	código del campus al que está asociado dicho departamento	
des_campus	nombre del campus al que está asociado dicho departamento	
cod_situacion_departamento	código de la situación del departamento a la fecha de referencia del dataset	<ul style="list-style-type: none"> • '1' para Activo • '2' para Inactivo • '3' para En proceso de extinción • '4' para Extinguido en el curso de referencia
des_situacion_departamento	nombre de la situación del departamento a la fecha de referencia del dataset	<ul style="list-style-type: none"> • 'Activo' • 'Inactivo' • 'En proceso de extinción' • 'Extinguido en el curso de referencia'
fecha_desde_situación_departamento	fecha desde la cual el centro se encuentra en la situación consignada	

Tabla 5. Campos del dataset de Departamentos [62]

5.3.4 Diagrama del software

El diagrama del software es muy sencillo para este conjunto de datos. Partimos de la página web de la Universidad Politécnica donde se muestran los campus y sus escuelas o facultades adscritas [37].

La aplicación, tras acceder a cada escuela, obtiene ciertos campos. Estos campos se manipulan para adecuarse a los estándares de Núcleo Común y se forma el dataset.



Figura 12. Diagrama del software para el dataset de Departamentos

5.3.5 Obtención de las variables

En este caso la aplicación es sencilla. Del enlace de cada escuela se recopila el campus al que pertenece y los departamentos que conforman dicha escuela.

- 📁 Campus Montegancedo
 - 📁 E.T.S. de Ingenieros Informáticos
 - 📁 Centro de Inv. en Biotecnología y Genómica de Plantas (CBGP)
 - 📁 Centro de Tecnología Biomédica (CTB)
 - 📁 Instituto de Microgravedad "Ignacio Da Riva"
 - 📁 Centro de Domótica Integral (CeDINT-UPM)

Figura 13. Campus Web UPM

En la parte superior derecha se muestran las variables constantes de este dataset.

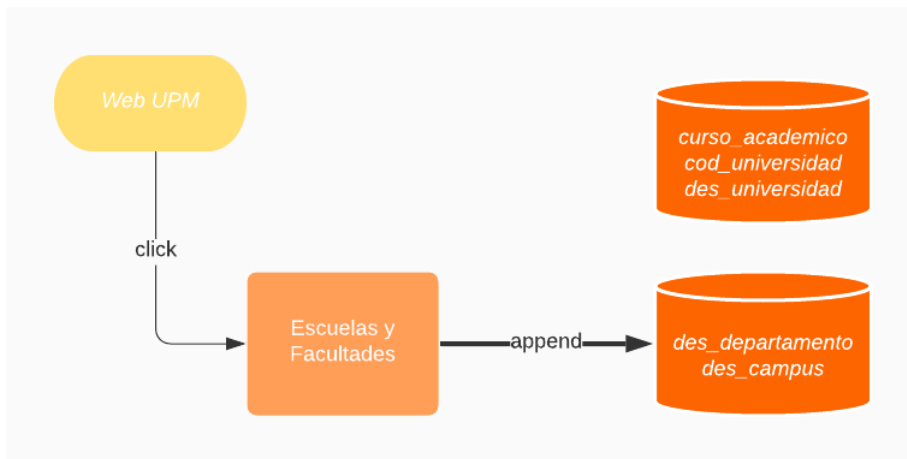


Figura 14. Obtención de variables del dataset de Departamentos

5.3.6 Consideraciones

Los siguientes campos no han podido ser obtenidos por unas circunstancias. Los campos en cuestión son:

- `des_municipio`: este campo no se ha obtenido por temas técnicos. No hemos sido capaces de obtener el municipio del mapa de Google Maps que hay en las páginas de cada centro.
- `cod_municipio`: si el campo '`des_municipio`' hubiese sido obtenido, este campo se obtiene de la misma forma que en el dataset anterior, utilizando el fichero descargado del INE.

5.3.7 Campos no obtenidos

Sin embargo, no todos los [campos](#) han podido ser obtenidos con las herramientas a las que tenemos acceso. En caso de que un campo no haya podido ser obtenido, su valor se reemplazará por un espacio en blanco ' '. En este caso los campos que no hemos podido obtener son los siguientes:

- `cod_departamento`
- `cod_municipio`
- `des_municipio`
- `cod_campus`
- `cod_situacion_departamento`
- `des_situacion_departamento`
- `fecha_desde_situacion_departamento`

En este dataset al recopilar los datos de una página web particular, no pertenece a ningún organismo público, la mayor parte de los campos correspondientes a códigos no han podido ser obtenidos.

5.3.8 Apariencia del código

En lo que respecta al código elaborado la estructura es la siguiente.

Primeramente, configuramos el controlador y accedemos a la página web. A continuación, accedemos a cada uno de los centros pertenecientes a cada uno de los campus (Figura 13).

```
driver_path = 'C:\\Users\\sanch\\Downloads\\chromedriver.exe'
option = Options()
option.add_argument('--headless') #ocultar pestaña navegador

#DEPARTAMENTOS
url = 'https://www.upm.es/UPM/Centros'
driver = webdriver.Chrome(driver_path, options=option)
driver.get(url)
```

Una vez estamos en la página específica del centro, obtenemos el nombre del centro y de la tabla de departamentos. Esta tabla la recorremos para obtener el cada uno de los departamentos.

```

centros_campus = driver.find_elements_by_xpath('//*[@id="subcana-
les"]/li[3]/ul/li')
links = len(driver.find_elements_by_xpath('//*[@id="sub-
canales"]/li[3]/ul/li/a'))

for centro in centros_campus:
    centros.append(centro.find_elements_by_xpath('..//a')[0].text)

for link in range(0, links):
    centro = driver.find_elements_by_xpath('//*[@id="subcana-
les"]/li[3]/ul/li/a')
    centro[link].click()
    campus.append(driver.find_elements_by_xpath('//*[@id="directo-
rio-03"]/div/h2')[0].text)

    departamentos_centro = driver.find_ele-
ments_by_xpath('//*[@id="directorio-03"]/div/div')
    departamentos_i = list()
    for depart in departamentos_centro:
        try:
            departamentos_i.append(depart.find_ele-
ments_by_xpath('..//a')[0].text)
        except:
            departamentos_i.append(depart.find_ele-
ments_by_xpath('..//div/div')[0].text)

    departamentos.append(departamentos_i)

```

Esto se ejecuta para cada uno de los campus. Posteriormente, tras unas modificaciones sobre los datos obtenidos inicialmente, obtenemos los valores finales.

5.4 Presupuesto de Ingresos dataset

5.4.1 Descripción

Este dataset contiene el presupuesto de ingresos de la Universidad Politécnica de Madrid. En él se recogen tanto la estructura del presupuesto como su situación inicial y el nivel de ejecución a una fecha de referencia [63]. Cada presupuesto está asociado a un año natural.

En cuanto a la estructura de presupuesto de ingresos de una administración pública, las partidas se estructuran en base a las denominadas clasificación económica y clasificación orgánica.

En la clasificación económica, el primer nivel de jerarquía son los capítulos que son un total de 9 [63]. Sin embargo, los dos primeros no se utilizan en el caso de las universidades:

1. Impuestos directos y cotizaciones sociales
2. Impuestos indirectos

3. Tasas, precios públicos y otros ingresos
4. Transferencias corrientes
5. Ingresos patrimoniales
6. Enajenación de inversiones reales
7. Transferencias de capital
8. Activos financieros
9. Pasivos financieros

En cuanto a los niveles de jerarquía para la clasificación económica quedarían de la siguiente forma [63]:

- Capítulo
 - Artículo
 - Concepto
 - Subconcepto
 - Partida

Por otro lado, la clasificación orgánica tiene otra distribución jerárquica, que se compone de 5 niveles [63]:

- Sección
 - Servicio
 - Unidad de coste
 - Orgánica libre 1
 - Orgánica libre 2

Para cada partida, se especificarán cuatro importes. Dichos importes representan la situación inicial prevista en el presupuesto, la situación actual a fecha de referencia y sus modificaciones, y el nivel de ejecución a fecha de referencia.

Este dataset está incluido en la categoría de 'Economía y Patrimonio'. Su frecuencia de actualización es anual. En el mes de febrero se publicará el presupuesto inicial del ejercicio en curso y en septiembre se publicará el presupuesto a cierre del ejercicio anterior. Para su diseño se han tenido en cuenta [63]:

- Ley 19/2013, de 9 de diciembre, de transparencia, acceso a la información pública y buen gobierno [15].
- Indicador 9.1 de la Fundación Compromiso y Transparencia. Una de las críticas más habituales a la universidad española es su falta de orientación a los resultados. Este indicador se asocia con la publicación de los resultados relacionados con la labor investigadora de sus profesores: tesis publicadas, publicaciones, patentes, etc [64].
- Indicador due421 de Dyntra [60]. Este indicador se relaciona con la publicación de los presupuestos de la universidad, incluyendo las principales partidas presupuestarias.

- Indicador due425 de Dyntra [60]. Este indicador está relacionado con la información trimestral o semestral sobre la ejecución presupuestaria.
- Indicador due426 de Dyntra [60]. Este indicador se asocia con el portal de Transparencia Económica-Financiera donde se muestran los datos del presupuesto y su estado de ejecución de forma gráfica, clara, accesible y entendible por la ciudadanía.

5.4.2 Fuentes de información

Este dataset ha sido formado a partir de la página web del Portal de Transparencia de la Universidad Politécnica de Madrid [38]. En el menú de 'Información Económica' podemos acceder a una serie de documentos informando sobre la procedencia y cuantía de los ingresos. Tras acceder a los presupuestos, los documentos utilizados para formar el presente conjunto de datos es el siguiente:

- Presupuesto Ingresos Clasificación orgánica [65]. Este documento en extensión XLS estructura los ingresos según la clasificación orgánica descrita previamente. El nivel de granularidad de este documento se encuentra en 'Servicio', aunque la mayor parte de los datos están referidos a la sección.
- Clasificación económica de ingresos por artículos [65]. Este documento en extensión XLS estructura los ingresos según la clasificación económica. En nivel de granularidad se sitúa en 'Subconcepto' según la clasificación económica.

5.4.3 Campos

Este dataset está compuesto por un total de 29 variables que suelen ir emparejadas por un código y su descripción.

Campo	Descripción
cod_universidad	código de la universidad
des_universidad	nombre de la universidad
anio	año natural al que se refiere el presupuesto
cod_capitulo	código del capítulo asociado al ingreso
des_capitulo	descripción del capítulo asociado al ingreso

cod_articulo	código del artículo asociado al ingreso
des_articulo	descripción del artículo asociado al ingreso
cod_concepto	código del concepto asociado al ingreso
des_concepto	descripción del concepto asociado al ingreso
cod_subconcepto	código del subconcepto asociado al ingreso
des_subconcepto	descripción del subconcepto asociado al ingreso
cod_partida	código de la partida asociada al ingreso
des_partida	descripción de la partida asociada al ingreso
cod_seccion	código de la sección asociada al ingreso
des_seccion	descripción de la sección asociada al ingreso
cod_servicio	código del servicio asociado al ingreso
des_servicio	descripción del servicio asociado al ingreso
cod_unidad_de_coste	código de la unidad de coste asociada al ingreso
des_unidad_de_coste	descripción de la unidad de coste asociada al ingreso
cod_organica_libre_1	código de la orgánica libre 1 asociada al ingreso
des_organica_libre_1	descripción de la orgánica libre 1 asociada al ingreso
cod_organica_libre_2	código de la orgánica libre 2 asociada al ingreso
des_organica_libre_2	descripción de la orgánica libre 2 asociada al ingreso

fecha_referencia	referencia temporal de la información económica consignada
credito_inicial	importe del crédito inicial
modificaciones	importe de las modificaciones
credito_total	importe del crédito total (crédito inicial + modificaciones)
derechos_reconocidos_netos	importe de los derechos reconocidos
recaudacion_neta	importe neto de los cobros

Tabla 6. Campos del dataset de Presupuesto de Ingresos [63]

5.4.4 Diagrama del software

Para la recopilación de datos se utilizan dos ficheros. Primeramente, se lee el documento 'Presupuesto Ingresos Clasificación Orgánica' que tiene un mayor nivel de granularidad al contener datos tanto de la clasificación económica como de la orgánica.

Posteriormente se carga el otro archivo. Ambos documentos se relacionan por el campo en común 'cod_subconcepto'. Este último documento describe la clasificación económica para cada uno de los ingresos.

Finalmente, los datos se guardan en un fichero CSV tras ser adecuados a los requisitos del Núcleo Común.

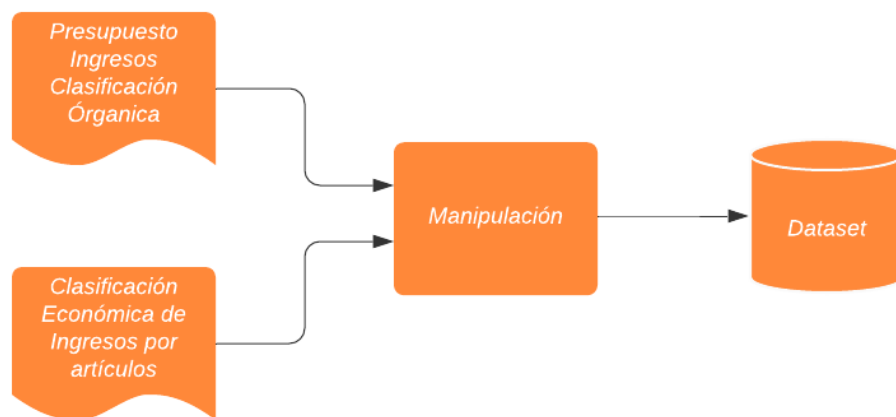


Figura 15. Diagrama del software para el dataset de Presupuesto de Ingresos

5.4.5 Obtención de las variables

En la siguiente imagen se puede visualizar los pasos seguidos para la recopilación de las variables, así como la fuente de información utilizada para obtener cada campo. En este dataset, la librería Pandas ha tenido un valor fundamental para el tratamiento de los datos.

El nexo de unión entre ambos datasets ha sido el campo relativo al código del subconcepto. Mediante este campo se relacionan cada una de los campos en sus respectivos datasets.

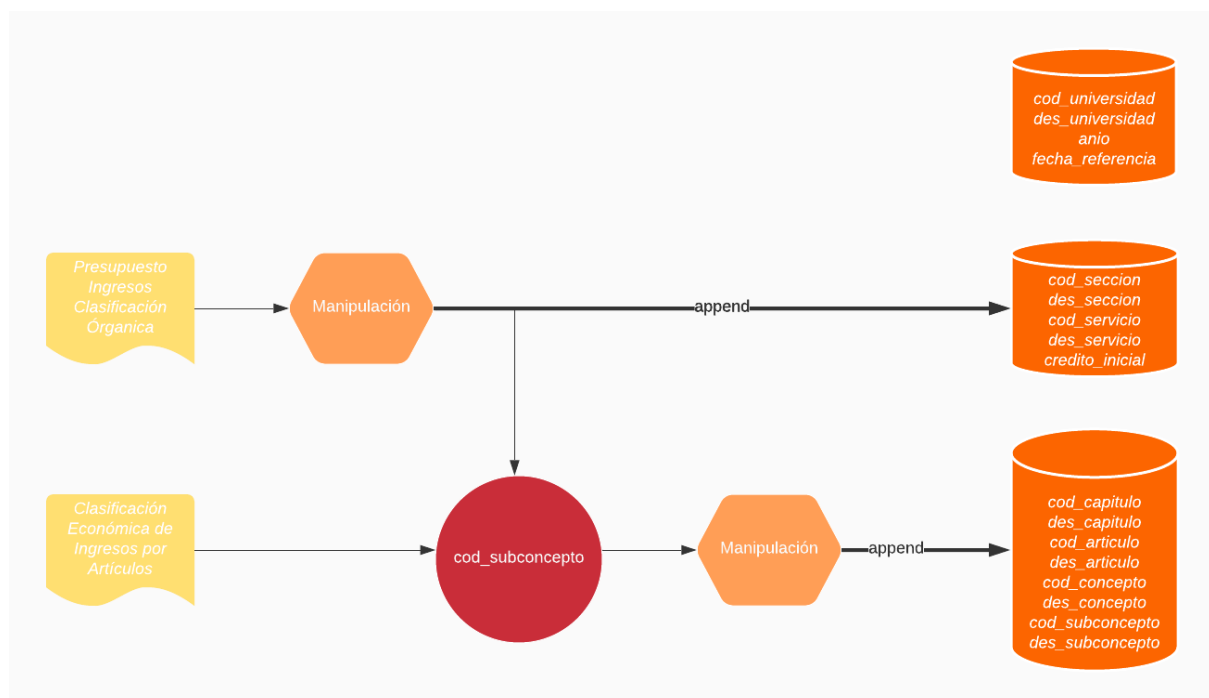


Figura 16. Obtención de campos para un ingreso

5.4.6 Consideraciones

Para este nos hemos encontrado con una serie de situaciones relaciones que vamos a proceder a aclarar:

- `fecha_referencia`: el valor de este corresponde al día en que se ejecuta la aplicación.
- `modificaciones`, `credito_total`, `derechos_reconocidos_netos`, `recaudacion_neta`: en este dataset en nivel de granularidad se sitúa en el campo 'Subconcepto'. Sin embargo, para estos importes en nivel de granularidad se sitúa en el campo 'Concepto' que es un nivel superior. El fichero se puede descargar en la pestaña 'Ejecución Presupuestaria' del menú 'Información Económica' con el nombre de 'Estado de ejecución del Presupuesto de Ingresos' [66]. Por esta razón para mantener la coherencia del documento se ha decidido mantener estos campos vacíos.

5.4.7 Campos no obtenidos

Sin embargo, no todos los [campos](#) han podido ser obtenidos con los recursos de los que disponemos. Estos campos se corresponden con los valores de mayor granularidad y por tanto que podrían contener información personal. Estos campos no obtenidos son los siguientes:

- `cod_partida`
- `des_partida`
- `cod_unidad_de_coste`
- `des_unidad_de_coste`
- `cod_organica_libre_1`
- `des_organica_libre_1`
- `cod_organica_libre_2`
- `des_organica_libre_2`

Además de estos campos, hemos decidido mantener vacíos los siguientes campos como comentamos en el apartado 'Consideraciones':

- `modificaciones`
- `credito_total`
- `derechos_reconocidos_netos`
- `recaudacion_neta`

5.3.8 Apariencia del código

En lo que respecta al código elaborado la estructura es la siguiente.

Por un lado, los valores obtenidos del primer fichero se obtienen de la siguiente forma.

```
for i in range(0, len(ingresos_organico)-3):
    if len(ingresos_organico.loc[i, 'CÓDIGO']) <= 2:
        seccion = ingresos_organico.loc[i, 'CÓDIGO']
        cod_seccion.append(ingresos_organico.loc[i, 'CÓDIGO'])
        des_seccion.append(ingresos_organico.loc[i, 'CENTRO'])
        cod_servicio.append('None')
        des_servicio.append('None')
    if len(ingresos_organico.loc[i, 'CÓDIGO']) >= 3:
        if '.' in ingresos_organico.loc[i, 'CÓDIGO']:
            cod_seccion.append(ingresos_organico.loc[i, 'CÓDI-
GO'].split('.')[0])
            des_seccion.append(ingresos_organico.loc[ingresos_or-
ganico['CÓDIGO'] == cod_seccion[1], 'CENTRO'][0])
            cod_servicio.append(ingresos_organ-
ico.loc[i, 'CÓDIGO'])
            des_servicio.append(ingresos_organico.loc[i, 'CEN-
TRO'])
        if '.' not in ingresos_organico.loc[i, 'CÓDIGO']:
            cod_seccion.append(ingresos_organico.loc[i, 'CÓDI-
GO'][:2])
            des_seccion.append(ingresos_organico.loc[ingresos_or-
ganico['CÓDIGO'] == cod_seccion[1], 'CENTRO'][0])
            cod_servicio.append(ingresos_organico.loc[i, 'CÓDI-
GO'])
            des_servicio.append(ingresos_organico.loc[i, 'CEN-
TRO'])

    afectado.append(ingresos_organico.loc[i, 'Afectado/No Afectado'])
    subconcepto.append(ingresos_organico.loc[i, 'CONCEPTO / SUB-
CONC.'])
    importe.append(ingresos_organico.loc[i, 'IMPORTE'])
```

En lo que respecta al otro documento, el código tiene cierta similitud con el mostrado anteriormente. Sin embargo, es más extenso y complejo puesto que tiene que obtener algunas de texto derivadas además de tratar distintos errores.

6. Conclusiones

En el presente trabajo se han desarrollado un total de 4 conjuntos de datos adecuados a las directrices del estándar [Núcleo Común](#). Para su obtención, el software ha sido creado con la mayor precisión posible para que este siga funcionando con el paso del tiempo, a pesar de que cambien ciertas estructuras de los árboles HTML de dónde se obtuvieron los datos.

Por otro lado, se anima a la comunidad de datos abiertos a reutilizar tanto el software como los contenidos generados con él. Además, se anima a cualquier desarrollador a mejorar el presente software con el fin de depurar errores o garantizar la correcta durabilidad del mismo. Este software se puede encontrar en el siguiente repositorio de GitHub [\[67\]](#).

Con respecto a las líneas futuras sobre las que se puede trabajar en el proyecto quedan enumeradas a continuación las que se plantean más interesantes:

- Solicitar acceso para obtener datos de carácter personal de utilidad para la formación de los datasets diseñados en Núcleo Común. Algunas de las fuentes que poseen este tipo de información son la UPM mediante su servicio [API](#) o mediante el [SIU](#).
- Desarrollar un portal open data que gestione nuestro catálogo de datos y nuestro contenido online. Este portal puede ser desarrollado mediante la herramienta DKAN [\[22\]](#). Esta herramienta es similar a CKAN, utilizada por ejemplo en la Iniciativa de datos abiertos del Gobierno de España [\[25\]](#), pero cuenta con ventajas como por ejemplo la gestión de noticias, presentación de reportes basados en datos, infografías y muchas más.

7. Bibliografía

- [1] “Instituto Nacional de Estadística.” [Online]. Available: <https://www.ine.es>
- [2] Educaweb, “El Espacio Europeo de Educación Superior (EEES).” [Online]. Available: <https://www.educaweb.com/contenidos/educativos/espacio-europeo-educacion-superior-eees/>
- [3] Boletín oficial del Estado, Legislación Consolidada, *Real Decreto 1393/2007, de 29 de octubre, por el que se establece la ordenación de las enseñanzas universitarias oficiales*. 2007. [Online]. Available: <https://www.boe.es/eli/es/rd/2007/10/29/1393/con>
- [4] Boletín oficial del Estado, *Real Decreto 861/2010, de 2 de julio, por el que se modifica el Real Decreto 1393/2007, de 29 de octubre, por el que se establece la ordenación de las enseñanzas universitarias oficiales*. Madrid, 2010. [Online]. Available: <https://www.boe.es/eli/es/rd/2010/07/02/861>
- [5] “Agencia Nacional de Evaluación de la Calidad y Acreditación - ANECA” [Online]. Available: <http://www.aneca.es/>
- [6] “Conferencia de Rectores de las Universidades Españolas - CRUE.” [Online]. Available: <https://www.crue.org/>
- [7] Conferencia de Rectores de las Universidades Españolas - CRUE, “La Universidad Española en Cifras.” [Online]. Available: <https://www.crue.org/publicacion/espanola-en-cifras/>
- [8] Fundación Conocimiento y Desarrollo (CYD), “El mayor ranking de universidades españolas.” [Online]. Available: <https://www.rankingcyd.org/>
- [9] U-Multirank, “World University Rankings.” [Online]. Available: <https://www.umultirank.org/>
- [10] BBVA Fundación y IVIE, “U-Ranking de las Universidades Españolas.” [Online]. Available: <https://www.u-ranking.es/index2.php#>
- [11] “Instituto Valenciano de Investigaciones Económicas - IVIE.” [Online]. Available: https://www.ivie.es/es_ES/
- [12] Universidad Excelente, “El Sistema Integrado de Información Universitaria (SIIU): una buena herramienta aún incompleta,” 2019. [Online]. Available: <https://universidadexcelente.com/el-sistema-integrado-de-informacion-universitaria-siiu-una-buena-herramienta/>
- [13] Ministerio de Universidades, “Estadísticas e Informes Universitarios.” [Online]. Available: <https://www.universidades.gob.es/portal/site/universidades/menuitem.78fe777017742d34e0acc310026041a0/?vgnnextoid=b93dd58bc3350710VgnVCM1000002006140aRCRD>
- [14] Gobierno de España, “Datos abiertos y reutilización de la información del sector público.” [Online]. Available: https://administracion.gob.es/pag_Home/espanaAdmon/Transparencia-y-datos-abiertos-2/datos_abiertos.html

- [15] Boletín oficial del Estado, *Ley 19/2013, de 9 de diciembre, de transparencia, acceso a la información pública y buen gobierno*. [Online]. Available: <https://www.boe.es/eli/es/l/2013/12/09/19/con>
- [16] Diario Oficial de la Unión Europea, “Directiva (UE) 2019/1024 del Parlamento Europeo y del Consejo relativa a los datos abiertos y la reutilización de la información del sector público (versión refundida)”, 2019. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=uriserv:OJ.L_.2019.172.01.0056.01.SPA
- [17] Gobierno de España, “Aprobada la reforma de la normativa europea sobre datos abiertos y reutilización de la información del sector público.” [Online]. Available: <https://datos.gob.es/es/noticia/aprobada-la-reforma-de-la-normativa-europea-sobre-datos-abiertos-y-reutilizacion-de-la>
- [18] UniversiDATA, “El portal de los DATOS abiertos sobre educación superior.” [Online]. Available: <https://www.universidata.es/>
- [19] “Dimetrical.” [Online]. Available: <https://www.dimetrical.es/es>
- [20] UniversiDATA, “¿Qué ventajas aporta a las Universidades publicar sus datos abiertos en UniversiDATA?” [Online]. Available: <https://www.universidata.es/productores>
- [21] Boletín oficial del Estado, *Ley 10/2019, de 10 de abril, de Transparencia y de Participación de la Comunidad de Madrid*. 2019. [Online]. Available: <https://www.boe.es/eli/es-md/l/2019/04/10/10/con>
- [22] DKAN, “DKAN Open Data Platform.” [Online]. Available: <https://getdkan.org/>
- [23] Open Fiscal Data Package, “Global Initiative for Fiscal Transparency.” [Online]. Available: <http://www.fiscaltransparency.net/ofdp/>
- [24] “Open Contracting Data Standard - OCDS.” [Online]. Available: <https://www.open-contracting.org/data-standard/?lang=es>
- [25] Gobierno de España, “Iniciativa de datos abiertos.” [Online]. Available: <https://datos.gob.es/>
- [26] “Open Spending.” [Online]. Available: <https://openspending.org/>
- [27] UniversiDATA, “La especificación ‘Núcleo Común.’” [Online]. Available: <https://www.universidata.es/la-especificaci%C3%B3n-n%C3%BAcleo-com%C3%BAn>
- [28] UniversiDATA, “Sobre el ‘Núcleo Común’”, 18-Jul-2020. [Online]. Available: <https://dimetrical.atlassian.net/wiki/spaces/UNC/pages/414122024/Sobre+el+N+cleo+Com+n>
- [29] Fundación Compromiso y Transparencia, “Rankings de Transparencia - Universidades.” [Online]. Available: <https://www.compromisoytransparencia.com/categorias-informes/universidades>
- [30] Dyntra, “Dynamic Transparency Index - Universities.” [Online]. Available: <https://www.dyntra.org/indices/universidades/>

- [31] E. T. Doñate and J. P. G. Mezquita, "Hacia una universidad abierta - Recomendaciones para el SUE." [Online]. Available: <https://tic.crue.org/wp-content/uploads/2016/03/Haciaunauniversidadabierta.pdf>
- [32] UniversiDATA, "Listado de Datasets," 24-Mar-2021. [Online]. Available: <https://dimetrical.atlassian.net/wiki/spaces/UNC/pages/413499396/Listado+de+Datasets>
- [33] UniversiDATA, "Directrices generales," 11-Nov-2020. [Online]. Available: <https://dimetrical.atlassian.net/wiki/spaces/UNC/pages/388300802/Directrices+generales>
- [34] Datatracker, "Common Format and MIME Type for Comma-Separated Values (CSV) Files." [Online]. Available: <https://datatracker.ietf.org/doc/html/rfc4180>
- [35] Ministerio de Educación y Formación Profesional, "Registro de Universidades, Centros y Títulos (RUCT)." [Online]. Available: <https://www.educacionyfp.gob.es/servicios-al-ciudadano/catalogo/estudiantes/donde-estudiar/ruct.html>
- [36] "Boletín oficial del Estado, *REAL DECRETO 1509/2008, de 12 de septiembre, por el que se regula el Registro de Universidades, Centros y Títulos*. 2008." [Online]. Available: <https://www.boe.es/buscar/doc.php?id=BOE-A-2008-15464>
- [37] Universidad Politécnica de Madrid, "Centros y Campus Universitarios." [Online]. Available: <https://www.upm.es/UPM/Centros>
- [38] Universidad Politécnica de Madrid, "Portal de Transparencia" [Online]. Available: <https://transparencia.upm.es>
- [39] Gobierno de España, "Administración General del Estado." [Online]. Available: https://administracion.gob.es/pag_Home/espanaAdmon/comoSeOrganizaEstado/Administracion_Gral_Estado.html
- [40] Instituto Nacional de Estadística, "Relación de municipios y sus códigos por provincias. Últimos datos." [Online]. Available: https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177031&menu=ultiDatos&idp=1254735976614
- [41] Dyntra, "La Plataforma Social de Transparencia y Gobierno Abierto." [Online]. Available: <https://www.dyntra.org/>
- [42] Selenium Documentación, "Selenium, el proyecto para automatizar navegadores." [Online]. Available: <https://www.selenium.dev/documentation/es/>
- [43] W3C, "WebDriver" [Online]. Available: <https://www.w3.org/TR/webdriver/>
- [44] "Google Chrome." [Online]. Available: <https://www.google.com/intl/es-es/chrome/>
- [45] "ChromeDriver - WebDriver for Chrome." [Online]. Available: <https://chromedriver.chromium.org/downloads>

- [46] "Python." [Online]. Available: <https://www.python.org/>
- [47] "Python 3.9.1." [Online]. Available: <https://www.python.org/downloads/release/python-391/>
- [48] "Sublime Text 3 - Download." [Online]. Available: <https://www.sublimetext.com/3>
- [49] Webdriver Documentación, "WebDriver." [Online]. Available: <https://www.selenium.dev/documentation/es/webdriver/>
- [50] Open Source Initiative, "Apache License, Version 2.0." [Online]. Available: <https://opensource.org/licenses/Apache-2.0>
- [51] "Pandas" [Online]. Available: <https://pandas.pydata.org/>
- [52] Open Source Initiative, "The 3-Clause BSD License ." [Online]. Available: <https://opensource.org/licenses/BSD-3-Clause>
- [53] "re - Regular expression operations." [Online]. Available: <https://docs.python.org/3/library/re.html>
- [54] Open Source Initiative, "The CNRI portion of the multi-part Python License (CNRI-Python)." [Online]. Available: <https://opensource.org/licenses/pythonpl.php>
- [55] "unicodedata — Unicode Database." [Online]. Available: <https://docs.python.org/3/library/unicodedata.html>
- [56] Open Source Initiative, "Python License (Python-2.0)." [Online]. Available: <https://opensource.org/licenses/Python-2.0>
- [57] Geopy, "Welcome to GeoPy's documentation!" [Online]. Available: <https://geopy.readthedocs.io/en/stable/>
- [58] Open Source Initiative, "The MIT License." [Online]. Available: <https://opensource.org/licenses/MIT>
- [59] UniversiDATA, "Dataset: Titulaciones," 12-Nov-2020. [Online]. Available: <https://dimetrical.atlassian.net/wiki/spaces/UNC/pages/417529859/Dataset+Titulaciones>
- [60] "Informe Dyntra sobre Universidades". Enero de 2019. [Online]. Available: https://www.dyntra.org/new/wp-content/uploads/2019/01/Informe_Universidades_Enero_2019.pdf
- [61] UniversiDATA, "Dataset: Centros," 12-Nov-2020. [Online]. Available: <https://dimetrical.atlassian.net/wiki/spaces/UNC/pages/417628163/Dataset+Centros>
- [62] UniversiDATA, "Dataset: Departamentos," 12-Nov-2020. [Online]. Available: <https://dimetrical.atlassian.net/wiki/spaces/UNC/pages/434798593/Dataset+Departamentos>

-
- [63] UniversiDATA, "Dataset: Presupuesto de Ingresos," 17-Nov-2020. [Online]. Available: <https://dimetrical.atlassian.net/wiki/spaces/UNC/pages/420642842/Dataset+Presupuesto+de+Ingresos>
- [64] J. M. Cavanna and E. Barrio, "Examen de transparencia - Informe de transparencia voluntaria en la web de las universidades españolas 2014," Oct. 2015. [Online]. Available: <https://gpaq.upc.edu/ranguings/doc/articles/Informe-transparencia-universidades-2014.pdf>
- [65] Portal de Transparencia UPM, "INFORMACIÓN ECONÓMICA Presupuestos." [Online]. Available: <https://transparencia.upm.es/economico/presupuestos>
- [66] Portal de Transparencia UPM, "INFORMACIÓN ECONÓMICA Ejecución Presupuestaria." [Online]. Available: https://transparencia.upm.es/economico/ejecucion_presupuestaria
- [67] Á. Sánchez, "TFM Open Data for Universities." [Online]. Available: <https://github.com/angeel2706/TFM-Open-Data-for-Universities>