Preprint of: Y. Tzitzikas, M. Pitikakis, G. Giakoumis, K. Varouha, E. Karkanaki,
What Process a University can Follow for Open Data? The University of Crete Case,
International Journal on Metadata Semantics and Ontologies (IJMSO), (accepted for publication), 2022

What Process a University can Follow for Open Data? The University of Crete Case

Yannis Tzitzikas

Computer Science Department, University of Crete
and Information Systems Laboratory, FORTH-ICS,
Heraklion, Greece,
E-mail: tzitzik@csd.uoc.gr

Marios Pitikakis

Computer Science Department, University of Crete,
Heraklion, Greece,
E-mail: pitikakis@csd.uoc.gr

Giorgos Giakoumis

Computer Science Department, University of Crete,
Heraklion, Greece,
E-mail: geokop7@gmail.com

Kalliopi Varouha

University of Crete,
Rethymnon, Greece,
E-mail: kvarouha@uoc.gr

Eleni Karkanaki

University of Crete,
Heraklion, Greece,
E-mail: eleni.karkanaki@uoc.gr

Abstract All public bodies in Greece, including Universities, are obliged to comply with the national legal framework and policy on open data. An emerging concern is how such a big and diverse organization could develop supporting procedures from an administrative, legal and technical stand point, that will enhance and expand the level of the provided open data-related services. In this paper, we describe our experience, as University of Crete, for tackling with these requirements. In particular, (a) we detail the steps of the process that we followed, (b) we show how an Open Data Catalog can be exploited also in the first steps of this process, (c) we describe the platform that we selected, how we organized the catalog and the metadata selection, (d) we describe extensions that were required, (e) we motivate and describe various additional services that we developed, and (f) we discuss the current status and possible next steps.

Keywords: Open Data; University Open Data; Data Sharing.

1 Introduction

Open access has been a core strategy in the European Commission for several years now and aims at improving knowledge circulation and thus innovation. In 2012, the European Commission encouraged all EU Member States, via a Recommendation, to put public-funded research results in the public sphere in order to make science more efficient and strengthen their knowledge-based economy.

Indeed, open data can help to accelerate knowledge and information sharing, which can lead to more efficient data management, reduce the replication or repetition of data collection, facilitate the develop information infrastructures, create more integrated services, and improve interactions between the private and public sector, among organizations or governments and people. Moreover, open data can increase transparency in the governments' decisions and make governments more efficient and more reliable. It is widely recognized that making data and research results more accessible contributes to

advancements in science and innovation in the public and private sectors. For example, the EU Open Data Portal[1] and European Data Portal[2] provide access to the European Union open data, categorized by subject and/or application domain.

In Greece, starting from October of 2014 with the law 4305 (amending the law 3448/2006) each public organization is obliged to comply with the Directive 2013/37/EU[3] of the European Parliament (amending Directive 2003/98/EC) on the *re-use of the public sector information* and this includes higher education area. In addition, in September of 2020 with law 4727 for Digital Government, the Directive 2019/1024 of the European Parliament and Council incorporates in the national law, in order to fully utilize information from public sector for the Greek economy and society. For the first time, special regulations are introduced in data categories of which Greece has a particularly high economic and social life benefit, such as dynamic data, research data and high value data.

Responding to the aforementioned requirements is a challenging endeavor for a university because (a) a university is a big organization, (b) it comprises schools, departments and units of different characteristics and mindset, (c) the legal framework should be fully respected.

The rising question is how a university could support open data administratively, legally and technically in a flexible way that allows for the gradual improvement of the open data provided? In this paper we describe what University of Crete has done so far for responding to the requirements of the national legal framework on open data. The key contributions of this (case study) paper are: (a) we detail the steps of the process that we followed (also from an administrative point of view), (b) we show how an Open Data Catalog can be exploited also in the first steps of this process, (c) we describe the catalog platform that we selected, how we organized the catalog and what metadata we included, (d) we describe extensions that were required, (e) we discuss the current status and possible next steps.

This is an extended version of [16]. With respect to that paper, the current paper:

- provides more good cases of university open data worldwide (§2.2),
- describes in more detail the situation in Greek Universities as regards open data (§2.3),
- provides more information about the governance or the U. of Crete open data (§5.1),
- motivates and describes various extra services that we developed over the catalog (§6)

The rest of this paper is organized as follows. Section 2 describes the general context and related work in national and international level, Section 3 describes the process that we followed and the envisioned ecosystem, Section 4 describes the Open Data Catalog, Section 5 describes supporting related activities, Section 6 describes extra services that were developed over the catalog, Section 7 discusses possible future steps, and finally, Section 8 concludes the paper.

2 Context and Related Work

2.1 General Context

Developing an open data strategy is not a trivial task. [5] discusses how to implement an open data strategy, but it focuses on the private sector, and suggests that open data requires substantial organizational change. [14] analyzes open government data usage from a social point of view. [12] discusses open data partnerships between firms and universities. [2] contains an interesting discussion and analysis of the challenges the universities face for stewarding the data they collect and hold in ways that balance accountability, transparency, and protection of privacy, academic freedom, and intellectual property. In general, the potential of open data is unlimited. Just indicatively [7] describes an approach for interlinking educational information across universities through the use of Linked Data principles, while [10] describes an approach for ranking universities using Linked Open Data. In the same context, of university rankings, [4] claims that an integration of existing data in an open-linked data platform may permit the construction of new indicators about universities, without having to design the indicators on a custom basis. Finally, [3] discusses the potential value of open data as a material for use in learning activities, while [13] presents a data-driven system to extract relevant information hidden in the student academic data and, thus, help tutors to offer their students a more proactive personal guidance.

2.2 Universities Worldwide

As regards international practice, below we list a few indicative universities that follow good practices as regards open data. The Harvard College Open Data Project (HODP)[4] is a student-faculty group that aims to increase transparency and solve problems on campus using public Harvard data, featuring dozens of publicly-available datasets from around Harvard University. The University of Southampton catalog[5] provides datasets in (at least) Turtle and RDF/XML formats, as well as an HTML description, and it also offers a SPARQL endpoint. The Pitt Data Catalog[6] is a tool to help researchers identify and locate datasets created by Univer-

[1] http://data.europa.eu/euodp/en/home
[2] https://www.europeandataportal.eu/
[3] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32013L0037&from=EN/
[4] https://hodp.org/
[5] https://data.southampton.ac.uk/datasets.html
[6] https://datacatalog.hsls.pitt.edu/

sity of Pittsburgh faculty and staff. There is an Open Data Initiative at the University of Waterloo (https://uwaterloo.ca/open-data/), which mainly provides an Open Data API as a service. Based on this API, students have developed a collection of several useful tools and applications (https://uwaterloo.xyz/). University of Oxford has an institutional linked open data store (https://data.ox.ac.uk/) where all data are also available through a SPARQL endpoint. Similarly, the UK Open University (https://data.open.ac.uk/) has a lot of datasets available through standard formats (RDF and SPARQL) which can be searched via a keyword-based interface or queried using the SPARQL endpoint. Aalto University (http://data.aalto.fi/) also has a Linked Data service and publishes datasets related to courses, publications, research projects, places, researchers and staff, organizational structure, news and events.

In general, we observe a gap in the literature as regards the processes that a university can follow as regards open data, i.e. in most cases the process is not described nor analyzed. One exception is [17] that describes the efforts of the University of Alicante (Spain), in the context of the OpenData4U project, in order to create a methodology to easily develop open data portals for universities.

2.3 National Level (Greek Universities).

In Greece there are 24 universities. 3 of them, namely, University of Crete[7] , Aristotle University of Thessaloniki[8] and University of Macedonia[9] have set up a catalog for open data. University of Macedonia has also selected CKAN platform to develop their catalog which contains 22 organizations and 132 datasets. Other universities have not set up a catalog, but have decided to directly publish their data in the central Greek Government portal for open data. Indeed, in that catalog, on July 2020, we found datasets from 10 universities. The number of datasets per university that were published was rather low, ranging from 3 to 14. The data published by the aforementioned (10 out of 24 Greek) universities, is summarized in Table 1. That table also describes the datasets that are published by each university. Currently, these datasets are hosted in http://archive.data.gov.gr/ since the central Greek Government portal (https://www.data.gov.gr/) for open data is currently (June 2021) in a transition phase; a beta version is online that provides access to just 47 series of data.

Overall, we could say that we are still in the first steps of the path of open academic data, as more than the half of Greek Universities have not published any dataset.

As regards, *data aggregation*, there is an aggregation infrastructure already established by the National Documentation Centre of Greece (EKT). EKT is responsible for the collection, development and maintenance of the National Archive of PhD Theses from all Higher Education Institutions in Greece which includes over 44K theses[10]. It also maintains the OpenArchive.gr[11], the largest portal providing a single point of access to Greek scientific digital content, produced by research, science and education bodies, with over 750K items from 59 institutions. It is also worth mentioning, that in June 2020, the "National Plan for Open Science" was released [1][12] focusing on open science at national level. Moreover, another research data repository[13] has started to develop by Hellenic Academic Libraries (Heal-link) but it is still at early stage since it publishes only two datasets so far.

3 The Process Followed and the Envisioned Ecosystem

Below we describe the five main steps of the overall process that UoC (University of Crete) has followed.

• $\mathcal{S}_{①}$: A *Task Force for Open Data*, for short TFOD, was appointed (June 2017) which prepared a feasibility study about Open Data.

• $\mathcal{S}_{②}$: The task force (TFOD) then performed an *inventory of the documents and datasets in the possession of the university*. For this purpose, an Open Data Catalog was installed and customized accordingly (the catalog is described in detail in §4).

• $\mathcal{S}_{③}$: The *ODEDIAD*, acronym in Greek of "Clarity and Open Data Project Management Team" was formally appointed (Dec 2018), in compliance with the circular DHD/F.40/2369 of 24 January 2017 from the Ministry of Administrative Reform and Electronic Governance. To enhance flexibility, a small scheme was decided, comprising a general coordinator, and coordinators for the *administrative* aspect, the *technical* aspect and the *open data catalog* aspect. Moreover, each department/unit of the university has appointed a contact person with whom ODEDIAD communicates. The composition of the team is described in §5.1. ODEDIAD then (a) *harmonized and characterized* the dataset descriptions collected in step $\mathcal{S}_{②}$, and prepared the *Decision on Open Data* that was formally approved by the University (April 18, 2019) and was submitted to the Ministry of Education and Religious Affairs.

• $\mathcal{S}_{④}$: The Open Data Catalog was improved and aligned with the decision (in $\mathcal{S}_{③}$), and the Departments/Units were encouraged to start uploading the actual datasets in the catalog.

• $\mathcal{S}_{⑤}$: The Open Data Catalog became public (Oct 18, 2019) and ODEDIAD is responsible for maintaining and improving the catalog.

[7] http://opendata.uoc.gr/
[8] https://opendata.auth.gr/
[9] http://data.dai.uom.gr/data/
[10] https://www.didaktorika.gr/eadd/?locale=en
[11] https://www.openarchives.gr/aggregator-openarchives/portal/?language=en
[12] https://www.athenarc.gr/en/news/live-now-joint-proposal-greek-research-and-technology-organizations-na
[13] https://hardmin.heal-link.gr/

Table 1 Ten (out of the 24) Greek Universities that publish Open Data

	Greek Institutes	Catalog	Datasets in data.gov.gr	# Datasets in data.gov.gr	Comments
1	School of Fine Arts		YES	3	The datasets cover lists of personnel, study guides, announcements, quality and financial data
2	Aristotle University of Thessaloniki	YES	YES	8	Catalog (https://opendata.auth.gr/) of 48 datasets developed with Swagger (an open source tool). The datasets cover admission and study statistics, teaching staff, studies programs, studies program courses, course information and university units
3	National and Kapodistrian University of Athens		YES	6	The datasets cover financial data, teaching staff statistics, undergraduate and postgraduate studies statistics, lists of personnel, scholarships and awards
4	Hellenic Open University		YES	5	The datasets cover library and information center online public access catalogue, link to university main page, undergraduate and postgraduate study guides
5	University of Ioannina (ex TEI of Epirus)		YES	6	The datasets cover financial data, external experts registry, internal evaluation reports, studies guide, erasmus program guide
6	University of Crete	YES	YES	1	Catalog (http://opendata.uoc.gr/) with 514 datasets developed with CKAN. The datasets coverage is described in §4.6
7	University of Macedonia	YES			Catalog (http://data.dai.uom.gr/data/) of 132 datasets developed with CKAN. The datasets cover student statistics, logos, undergraduate and postgraduate programs guides, erasmus information, teaching staff, regulations. The datasets are not categorized so it is difficult to locate the desired information
8	University of Western Macedonia		YES	4	The datasets cover student statistics, evaluation reports, financial data and information for university units
9	University of West Attica (ex TEI of Athens)		YES	6	The datasets cover postgraduate study programs, financial data for research, financial and quality data, personnel's phone directory
10	International Hellenic University (ex TEI of Central Macedonia)		YES	14	The datasets cover financial data for research, payrol statistics, financial data, internal and external evaluation reports, internal and external experts registry, link to institutional repository, library and information center online public access catalogue, studies guide, personnel's phone directory

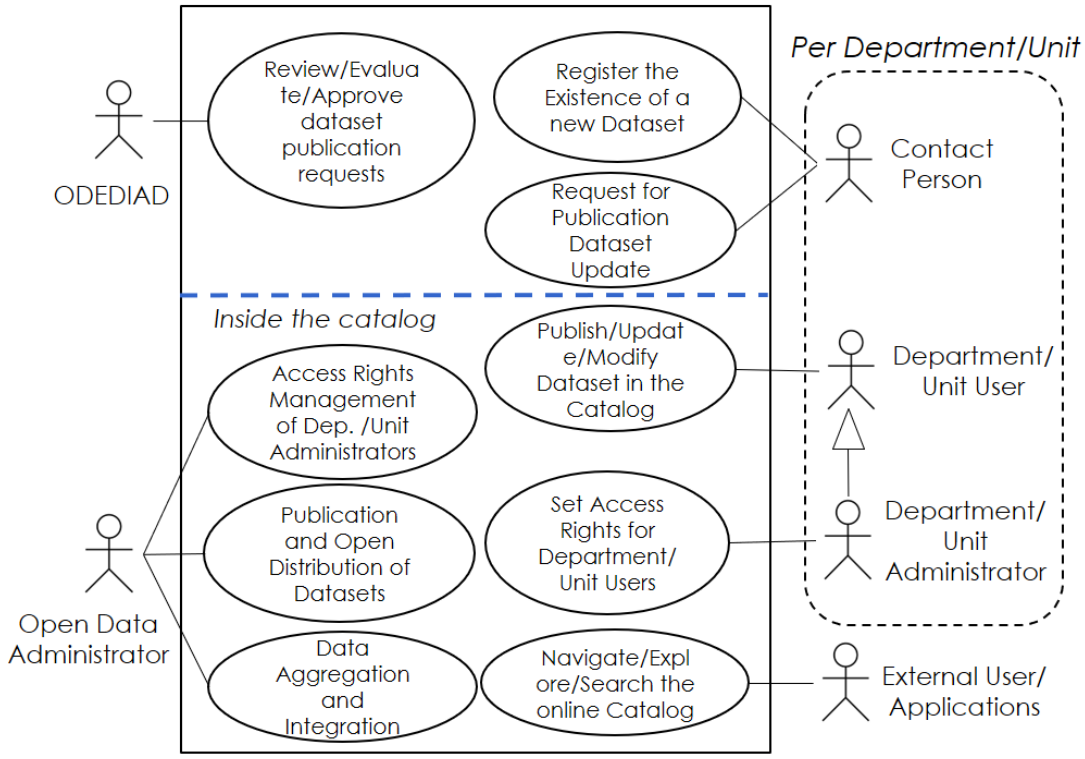

Figure 1 The Envisioned Ecosystem

It is worth noting that the process started in a bottom-up manner (at step $\mathcal{S}_{②}$), and then we applied a top-down harmonization step (at Step $\mathcal{S}_{③}$).

The envisioned ecosystem is shown in Figure 1 in the form of a Use Case Diagram that depicts the main Actors and the main Use Cases. The role of catalog-related actors is discussed in §4.3.

4 The UoC Open Data Catalog

Here we describe: how the platform was selected (in §4.1), how we customized CKAN (in §4.2), how we organized the catalog in department/units (in §4.3), what metadata we selected to include (in §4.4), what access services are offerred (in §4.5), and an analysis of the uploaded content so far (in §4.6).

4.1 Selecting the Platform

For selecting the University of Crete Open Data Catalog we investigated several options based on the platform characteristics, the technology considerations and the options that were available (Oct 2017). The cost of data catalogs is a key evaluation criterion. Products distributed as "open source" software are generally preferred because they are essentially "free" and can be modified or customized without restriction or licensing fees. However, open source software still requires management costs (hosting, maintenance, updates and security patches, training etc). On the other hand, SaaS products typically are proprietary; a vendor provides the software, the setup and hosting services, at a monthly or annual fee. Under the SaaS delivery model, the vendor is responsible for maintenance, server availability and reliability, scalability and performance according to a contract.

For making our decision we considered the following:
a) Self-managed, open source catalogs can provide a high degree of customization and autonomy. Most open source catalogs are designed to run in combination with other open source software and therefore technical proficiency in these areas is required.
b) The open data catalog must be hosted on reliable and relatively fast server architecture. Slow response times or periods of unavailability will discourage users.
c) Institution policies and/or laws regarding the open data catalog.
d) Scalability. As datasets are added, the catalog must be able to handle the additional load and must be easily extensible (have the flexibility to include additional functionality).
e) How datasets are managed and stored in the data catalog: "All-in-One" vs "Federated" catalogs. In an "All-in-One" model, datasets are stored in the catalog's architecture, with the main benefit of hosting and managing from a single platform and thus exercising strong oversight over the entire catalog infrastructure. Alternatively, in "Federated" catalogs datasets can reside on any publicly accessible location and the catalog includes a link (URL) to the dataset, as opposed to including the dataset itself.

f) Communities of support, support models or user communities.

Below we list and summarize the leading open data platforms.

CKAN[14] (Comprehensive Knowledge Archive Network) is a web-based open source management system for the storage and distribution of open data, formally supported by Open Knowledge International[15]. CKAN can be installed on any Linux server, including cloud-hosted configurations. CKAN is written in the Python programming language, and is designed for publishing and managing data either through a user interface or through an API. CKAN has a modular architecture through which additional or custom features may be added. Many national governments (like the United Kingdom, the United States, Austria, Brazil, Canada, Germany, Netherlands, Norway etc.) use CKAN to power their open data portals and it is also used by international portals such as the EU Open Data Portal[16]. Half of the founding members of the Open Government Partnership[17] use CKAN (as do four of the top five countries in the Global Open Data Index[18], and three of the top five countries in the Open Data Barometer[19]).

DKAN[20] is also open-source and takes a different approach by integrating the open data catalog features of CKAN into a content management system (CMS), namely Drupal. The underlying API is identical to CKAN but is based on Drupal, a popular content management system written in PHP instead of Python. This can be more appealing to organizations that have already invested in Drupal-based websites since DKAN can also be enabled in existing Drupal sites so that anyone can easily start to publish open data in standards compliant ways. Drupal has its own modular architecture with thousands of modules available for download (custom modules are also possible), and a large developer community.

Junar[21] delivers a cloud-based open data platform and manages its content based on the Software-as-a-Service (SaaS) model. This platform is frequently selected because of its ease of deployment and because it provides an "All-in-One" infrastructure. Junar can either provide a complete data catalog or can provide data via an API to a separate user catalog.

JKAN[22] is using Jekyll[1] (a simple static site generator) and allows for a quick deployment of static pages from underlying files. This data portal is based on CKAN and it is aimed at data publishers that need to deploy their data quickly. It is also open source, light-weight and can be easily customized with themes.

The Open Government Platform[23] (OGPL), like DKAN, is an open-source Drupal-based data catalog, but not designed to be CKAN-compatible at the API level. OGPL was jointly developed by the Government of India and the U.S. Government.

Socrata[24] is a cloud-based SaaS open data catalog platform that provides API, catalog, and data manipulation tools. One distinctive feature of Socrata is that it allows users to create views and visualizations based on published data and save those for others to use. Additionally, Socrata is proprietary but offers an open source version of their API, intended to facilitate transitions for customers that decide to migrate away from the SaaS model.

Our review revealed that the format in which metadata are published depends highly on the open data portal that publishes it. Open data portal software frameworks are either built on their own standards or use an already existing standard. The two predominant platforms – CKAN and Socrata – have each been developed on their own respective frameworks that are coming from major standards such as Dublin Core[25] and RDFS[26]. The platforms tend to use either one standard to generate metadata or a combination of a few. The most commonly used metadata standards in Open Data catalogs are based on some version of the Data Catalog Vocabulary (DCAT[27]) standard. DCAT has gained popularity due to its flexibility and elegant design, and aims at improving the data catalogues' interoperability so applications can easily consume metadata (even from multiple catalogs).

CKAN stores the datasets as a folder that hosts datasets or resources. The metadata are served as RDF and the platform supports DCAT, Dublin Core and INSPIRE[28] format. DKAN metadata fields are compatible with CKAN. Metadata are presented at a dataset level using Dublin Core, DCAT and INSPIRE geospatial format. Other standards setting organizations include the US FGDC and the European INSPIRE Metadata Directive. Junar uses the RDF metadata standard as presented in Dublin Core and DCAT, it does not support structural metadata. JKAN is based on CKAN, therefore it supports the same metadata standards as previously mentioned for CKAN. Socrata is based on the RDF metadata (Dublin Core and DCAT) with enrichment from custom metadata fields.

4.2 Adopting and Customizing CKAN

For the UoC Open Data Catalog we have selected CKAN to publish, store and manage open datasets. A dataset is a "unit" of data and contains two things: *metadata* (i.e. information about the data) and any number of *resources*, which hold the data itself. CKAN can store a

[14] http://ckan.org
[15] http://okfn.org
[16] https://open-data.europa.eu/
[17] https://www.opengovpartnership.org/
[18] http://index.okfn.org/
[19] http://https://opendatabarometer.org/
[20] https://getdkan.org/
[21] http://www.junar.com
[22] https://jkan.io/
[23] http://ogpl.github.io/index-en.html
[24] https://www.tylertech.com/products/socrata
[25] http://dublincore.org/documents/dces/
[26] https://www.w3.org/TR/rdf-schema/
[27] https://www.w3.org/TR/vocab-dcat/
[28] https://inspire.ec.europa.eu/data-specifications/2892

resource internally, or store it simply as a link (i.e. the resource itself could be elsewhere on the web). CKAN allows the creation of custom metadata fields, supports Dublin Core/DCAT for publishing metadata and provides data in open, non-proprietary formats such as CSV, XML, and JSON.

CKAN provides a *full-text search* for dataset metadata as well as filtering and sorting of results. It is possible to restrict the search to datasets with particular tags, data formats etc or targeted search within an organization/department. When a dataset is found and selected, CKAN displays the dataset page, which includes the name, description, and other information about the dataset, and links to and brief descriptions of each of the resources that belong to the dataset. On-site data/resource preview is also available for known data formats (like csv, xls, xlsx, rdf, xml, rdf+xml, owl+xml, atom, rss, json, geojson, png, jpeg, gif)

CKAN offers flexibility to integrate or embed the data catalog with other websites, add additional pages, layouts, color schemes, logos and generally easy customization. It has a large community of users and support, and offers extensibility with many additional or custom features developed regularly via extensions. CKAN also provides multi-language support and internationalization.

Plug-ins. CKAN provides a DCAT extension/plug-in[29] to export and harvest RDF serializations of datasets based on DCAT. The Data Catalog Vocabulary (DCAT) is "an RDF vocabulary designed to facilitate interoperability between data catalogs published on the Web". The extension defines the mapping of metadata for CKAN datasets and resources to the corresponding DCAT classes, mainly `dcat:Dataset` and `dcat:Distribution`, which is compatible with DCAT-AP v1.1[30]. DCAT-AP was designed to meet the metadata publishing needs in the context of the European Commission's Interoperability Solutions for European Public Administrations (ISA) programme: "Improving semantic interoperability in European eGovernment systems". This standard uses the main classes defined by the DCAT standard, and builds on properties to expand the fundamental version of the standard to fit the needs defined by the European Commission. This DCAT extension offers the following features: RDF DCAT Endpoints that expose the catalog's datasets in different RDF serializations (i.e. RDF/XML, Turtle, Notation3, JSON-LD), a RDF harvester that allows importing RDF serializations from other catalogs to create CKAN datasets, a JSON DCAT harvester that allows importing JSON objects that are based on DCAT terms.

Other CKAN extensions that are used for the UoC Open Data catalog are:

- A geospatial viewer[31] for CKAN resources, which contains view plugins to display geospatial files and services in CKAN. More specifically, we are using the Leaflet GeoJSON viewer plugin.
- The pages extension[32] that provides an easy way to add simple pages to CKAN. Using this extension we added a "featured datasets" page and Twitter feed page (for embedding Twitter content, news, dissemination activities etc)
- An LDAP plugin[33] which provides LDAP authentication for CKAN and integration with our existing UoC LDAP user service. This extension allows us to import existing username/full name/email/description, add LDAP users to a given organization automatically or combine LDAP and basic CKAN authentication.
- An extension[34] that integrates Google Analytics into CKAN.

In addition, some custom made extensions were created for the purposes of our UoC Open Data Catalog. For example, normally each dataset is owned by an "Organization" in a CKAN instance. A plugin was created to rename the default "Organization" of CKAN into "Department". Another plugin was developed to easily export the list of all datasets from a "Department". Some required additional metadata fields were inserted with another plugin and finally the appearance/theme of the catalog was customized.

4.3 Organization of the Catalog

There are different ways to categorize datasets related to University activities. The most important categories include: basic information about the University (e.g. the organizational structure, Schools and Departments, size, location, budget etc.), campus information (geographic information, maps, buildings etc.), educational administrative information (like courses, study programs, exams, curriculum, educational facilities, schedule etc.), staff and department information (professors and researchers, their basic information, contact details, research-related information etc.) student information (admissions, undergraduate/postgraduate statistics etc.).

Dataset cataloguing was performed in a bottom-up manner. We first collected all datasets by inviting contact persons to use the installed CKAN platform and to insert their datasets and related metadata (i.e. the descriptions of their datasets) instead of filling in spreadsheets. This approach contributed to a rapid conclusion of the cataloguing process, thus saving a lot of valuable time.

Currently (July 2021), there are 47 "departments" (16 academic, 22 administrative and 9 other) inserted into the catalog that correspond to the actual structure of the University of Crete. We have created another "department" for the UoC OpenData Task Force, which contains

[29] https://github.com/ckan/ckanext-dcat
[30] https://joinup.ec.europa.eu/release/dcat-ap-v11
[31] https://github.com/ckan/ckanext-geoview
[32] https://github.com/ckan/ckanext-pages
[33] https://github.com/NaturalHistoryMuseum/ckanext-ldap
[34] https://github.com/ckan/ckanext-googleanalytics

instructions and reference material shared between the OpenData team, the dataset administrators and the catalog users.

Each catalog "department" has its own administrator and multiple members (which can be either editors or simple users). Depending on each user's role in the department, users can perform different actions. A department administrator can edit the department's information, add/edit a dataset or add individual users to the department, with different roles depending on the level of authorization needed. An editor can create a dataset owned by that department. By default, a new dataset is initially private, and visible only to other users in the same department. When it is ready for publication, it can be published by the department administrator (this may require a higher authorization level within the department). A simple user in a department can only view private (and public) datasets.

During the $\mathcal{S}_{②}$ inventory phase, all users (contact persons, local administrators, department/unit users) were allowed to insert datasets. The uploaded datasets/resources were reviewed by the ODEDIAD and the University DPO in order to be compliant with the data protection directive. After phase $\mathcal{S}_{⑤}$, when the catalog became public, new datasets/resources can still be uploaded to the catalog but need to be reviewed, evaluated and approved by ODEDIAD and the DPO before they become available for public consumption and use. In addition, there is an annual validation and update procedure of all the UoC open datasets.

4.4 Selected Metadata

The metadata fields (per category) for each set of documents, information and data are shown in Table 2. They are categorized to: (a) general ones (title, quantity, etc), (b) rights/licenses-related, (c) metadata for digital files/datasets, (d) metadata for physical files, and (e) topic-related.

4.5 Access Services

In brief, the user can *search* the catalog, can *browse* the catalog by department, and can *explore* in a faceted-search manner the available datasets (through tags, file formats, access rights) as shown in Fig. 3 (right). Apart from the above *programmatic access* is supported. The native CKAN API allows users, providers, consumers and developers to write code that interacts with the CKAN site and its data. CKAN's Action API is a RPC-style API that exposes all of CKAN's core features to API clients. All of a CKAN website's core functionality (everything you can do with the web interface and more) can be used by external code that calls the CKAN API.

For example, using the CKAN API an application can: (a) Search/query data catalog, i.e. get lists of the site datasets, groups or other CKAN objects matching a specific query, (b) API access in multiple formats e.g. get a CSV, XML or JSON-formatted representation of a dataset, (c) Dataset/resource/object creation, update or deletion via the API (for data providers), and (d) Update metadata via the API (for data providers). More information and details on using the native CKAN's API can be found here.

There are some efforts to provide a Java API for accessing CKAN catalogs. Jackan, a Java client library for CKAN catalogs, is the most prominent one. It allows reading from and writing to CKAN catalogs, searching and filtering datasets, JSON serialization/deserialization, convert CKAN objects to their DCAT representation etc. Another effort is ckan4j, a CKAN extension Java client library that also provides access to the CKAN core functionalities and APIs. The difference with respect to other Java libraries for CKAN is the extension architecture which is designed not only to act as a Java client of a CKAN but also to extend the CKAN API adding specific enterprise grade functionalities like: Social Login (supported GitHub, Google+ and Facebook), CKAN dataset and organizations statistics exposed as API, CKAN tags statistics and classification data dump in JSON structures ready for D3.js visualizations etc.

Additional access services that are not supported by CKAN but we developed them on top afterwards will be described in §6.

4.6 Analysis of the Uploaded Contents

After several rounds of discussions with all the academic department contact persons (during $\mathcal{S}_{③}$), we arrived to a commonly agreed upon collection of datasets (homogenization/harmonization) that every department should provide openly (and preferably using in a machine processable format). This collection includes the 22 datasets for the 16 Departments of UoC that are shown in Table 3. Similarly, for the 5 Schools of UoC, a 5 list of common datasets were identified.

- Decisions of the bodies of the School
- Various admission and study statistics in the School
- Elections of the School collective bodies
- Correspondence and replies of administrative authorities
- School budget execution data (procurement process, invoices, etc.)

The rest of the University units had unique datasets and could not be classified in a uniform manner.

The Open Data Decision of the University of Crete, that contains the harmonized descriptions of all datasets of the organization is publicly available[35]. As mentioned in §3 (Step $\mathcal{S}_{④}$), following the issuance of this

[35] https://www.uoc.gr/files/items/7/7133/407_aps_sxed.1_odediad_omix469b7g-ayb.pdf

Table 2 Metadata fields: General, Rights-related, for Digital Files, for Physical Files

General fields	
1	Title
2	Description
3	File Owner / Location (e.g. Office, Department, Address)
4	Quantity of data (indicative) e.g. number, size in GB, number of documents etc.
5	Refresh rate (approx.)
6	Is it geospatial data? (YES/NO)
Rights of use and licences	
7	Availability (YES/NO)
8	If already available, way of accessibility (e.g. website, on request etc.)
9	If available through on request (either electronic or printed application) method of accessibility (e.g. by mail, in person, electronically etc.)
10	Is there a privacy/personal data restriction? (YES/NO)
11	Are the other restrictions? (e.g. national security issues, tax secrecy etc.) (YES/NO)
12	Available through fees? (YES/NO)
13	Available through licensing? (YES / NO / If YES, type of licence)
14	File Type (Physical/ Digital / Physical and Digital)
For digital file/dataset	
15	Automatically machine processable & editable format (e.g. txt, csv, html, xml, rdf, odt, ods, doc, xls, docx, xlsx)
16	Non automatically processable format (e.g. jpeg, tiff, gif, pdf, scanned documents etc)
17	Access through URL (e.g. website if available)
18	API (if available)
For physical file/dataset	
19	Format (e.g. A4 document, photographic archive etc)
20	Location (e.g. file cabinet 5 in room 123)
Additional information	
21	Content topic or tag (e.g. health, environment, economy, specific science etc.)

Table 3 The Datasets of an Academic Deparment

1	Open Courses	12	Events, Summer Schools, Meetings, Conferences, Competitions
2	Course Outlines	13	Photo Gallery
3	Undergraduate and Postgraduate Study Guides	14	Internship Programs
4	Course Schedules	15	Alumni information (after explicit consent of the graduates)
5	Admission and Study Statistics	16	Announcements and News
6	List of Publications of the Teaching, Scientific and Research staff	17	List of Personnel
7	MSc and PhD theses	18	Elections of collective bodies
8	Internal/external Evaluations and Accreditation	19	Financial information and data
9	Distinctions and Achievements	20	Decisions on committees establishment
10	Building facilities, maps, infrastructure	21	Decisions related to studies
11	Promotional/Dissemination material	22	Decisions on Tender Notices

decision, the Open Data Catalog was improved and aligned with that decision. Subsequently, the Departments/Units were encouraged to start uploading the actual datasets in the catalog or add links to the related web resources in case they are available in other systems. Currently (June 2021), the UoC Open Data Catalog[36] hosts a total of 509 descriptions of datasets, and 104 (20,76%) of them contain at least one resource or external link to a web page or institutional system. In machine processable formats (CSV, JSON, GeoJSON, XLS), we have 9 (1,8%) datasets. Note that some datasets are offered in more than one formats. Some especially useful datasets, are those that provide statistics about the UoC student population from 2006 up to now (gender, age, nationality, duration of studies etc.). Another useful dataset, is the "Research Directory" that contains information about all labs of the University, categorized by research field, department, position, or name. Finally, the Computer Science Deparment has posted many datasets in CSV and JSON format. Currently, the number of *tags* inserted into the CKAN catalog are 118.

5 Supporting Activities

Here we describe supporting activities, in particular §5.1 describes governance, §5.2 describes training and outreach activities, and §5.3 describes key performance indicators.

[36] http://opendata.uoc.gr/

5.1 Governance

As mentioned in §3, step $\mathcal{S}_{③}$, ODEDIAD (acronym in Greek of "Clarity and Open Data Project Management Team") was appointed. To enhance flexibility, a small scheme was decided, comprising a general coordinator, and coordinators for the *administrative* aspect, the *technical* aspect and the *open data catalog* aspect. Moreover, each department/unit of the university has appointed a contact person with whom ODEDIAD communicates. The organizational chart is given in Figure 2. The team meets regularly, every two months.

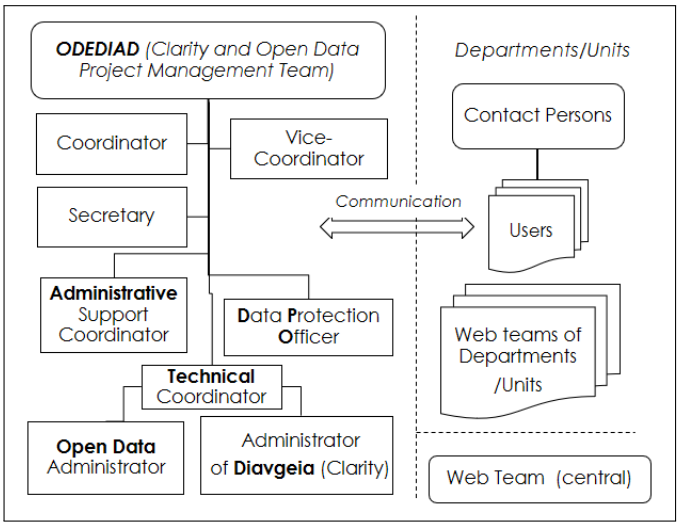

Figure 2 Governance of Open Data

5.2 Training and Outreach Activities

A manual with examples of how to use the catalog was created and published in the unit of ODEDIAD in the catalog (also for the needs of the step $\mathcal{S}_{②}$ as described in §3).

The catalog became public on Oct 18, 2020 (Figure 3 show the first page of the UoC Open Data catalog website) and all members of the university were informed through email and public announcements on the UoC website. Moreover the catalog was presented publicly in a wider audience, in a Conference on Open Data and GDPR, organized by by the Training Institute of the National Center for Public Administration and Local Government (EKDDA), 16 Dec 2019, Heraklion, Greece. To support also notification services to those interested in the updates of the catalog, a twitter account[37] was created on April 2020. The engagement of the community is rather low, however we have to note that no outreach action took place during the covid-19 pandemic period.

5.3 Monitoring and KPIs

We decided a number *Key Performance Indicators* (KPIs) for measuring and monitoring the performance. Their definition, and 2021 value, are described next.

- K1: Number of Datasets in the Open Data Decision of the University of Crete (June 2021 value: 501).
- K2: Number (and Percentage) of Datasets in the Open Data Catalog with at least one resource or external link to a web page or institutional system (June 2021 value: 104 (20,76%)).
- K3: Number (and percentage) of Datasets in the Catalog available in machine-processable form (June 2021: 9 (1,8%)).
- K4: Number of Visits to the Catalog (we started counting the visitors on April 2020). (June 2021: 475 (12% of them are returning visitors)).

Our short term goal is to increase K2 and K3.

6 Additional Services

We realized that more services are required for better exploiting the contents of the catalog now and in the future. This concerns *search* (described in §6.1), REST APIs to important datasets (described in §6.2), and import and export services for connectivity with external systems (described in §6.3). Figure 4 shows the overall architecture that includes these additional services.

6.1 Advancing the Search Service of the Catalog

One important requirement that we identified was to advance and extend the *search* functionalities since the offered search service by CKAN can only be applied to the metadata of the datasets. Although this can be useful for locating a dataset, it does not allow searching inside datasets e.g. for finding one particular name, laboratory, course, etc.

For this reason we decided to provide keyword search services (full text search) over the *contents* of the datasets. We utilized Elasticsearch[38] for indexing all text documents in our CKAN Open Data Catalog (e.g. pdfs, docs etc.). The user interacts with the system through a graphical interface, enters a keyword in the search bar and as a result all documents containing the specific keyword are displayed (see Figure 5).

The search application consists of three independent parts:

(a) the client (frontend) is an Angular application that consumes the API provided by the backend and displays the search results

(b) the Java-based backend application that implements the API used by the client

(c) the Elasticsearch (data persistence and search engine) which, after the text analysis and indexing procedure, stores indexes in JSON format (one for each different language identified) and accepts search requests and responds accordingly.

[37] https://twitter.com/UoC_OpenData

[38] https://www.elastic.co/elasticsearch/

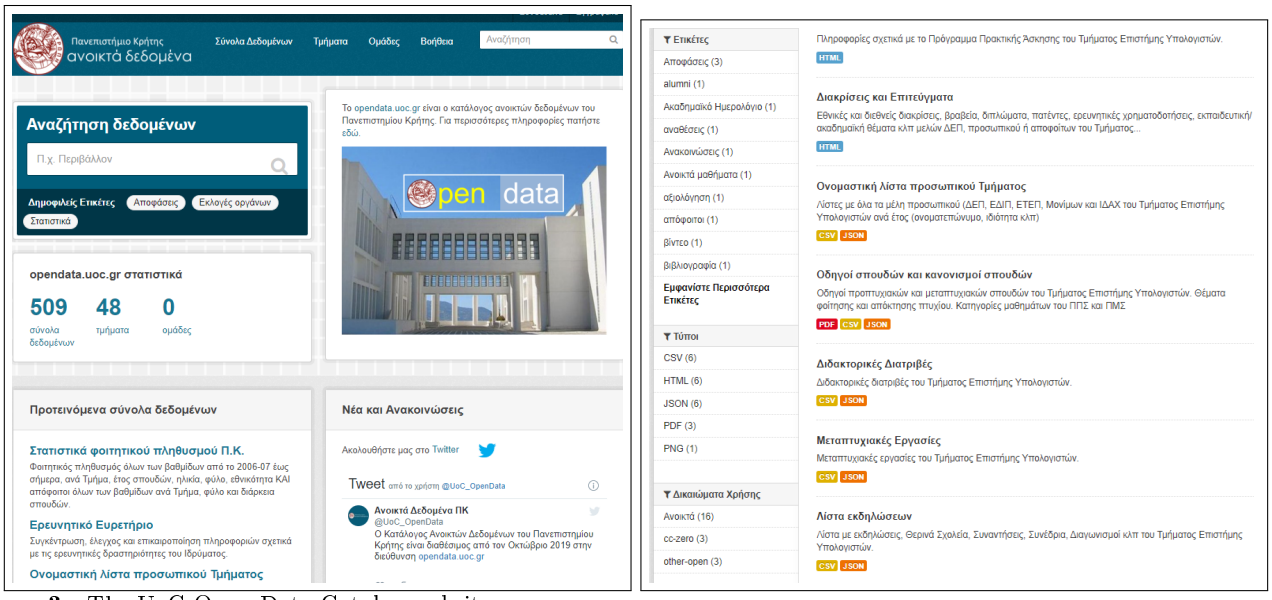

Figure 3 The UoC Open Data Catalog website

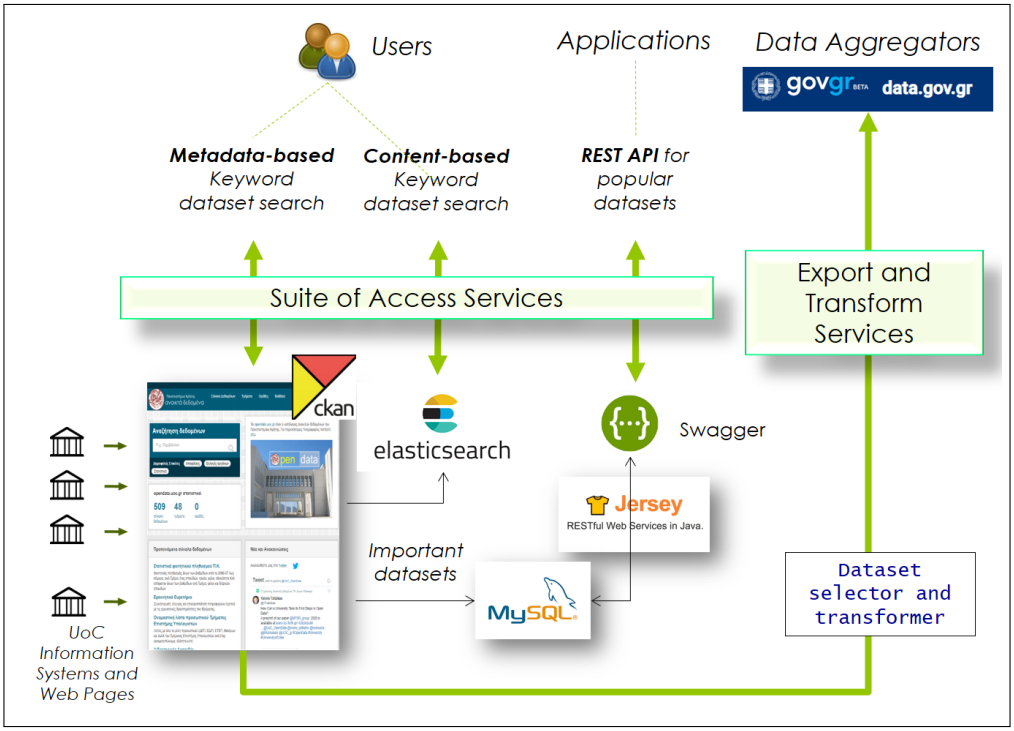

Figure 4 The Overall Architecture and includes the additional services

The application periodically synchronizes all new or modified documents from the UoC Open Data Catalog to the Elasticsearch and updates the indexes.

6.2 REST API and Clients

REST Web Services are frequently used to develop resource-oriented architectures. Therefore, we designed and developed REST style APIs for exposing web services related to some of our *most popular* datasets. We selected the UoC student population historical data, which consist of the following datasets (http://opendata.uoc.gr/el/dataset/ooithtikoe-ttahoyemoe-tt-k-2006-16):

1. Undergraduate students by gender and year of study.
2. Undergraduate students by gender and age.
3. Postgraduate students by year of graduation and gender.
4. Graduate undergraduate students per year of graduation.
5. Postgraduate students by year of study and gender.
6. PhD candidates by year of study and gender.
7. PhDs per year of graduation and gender.

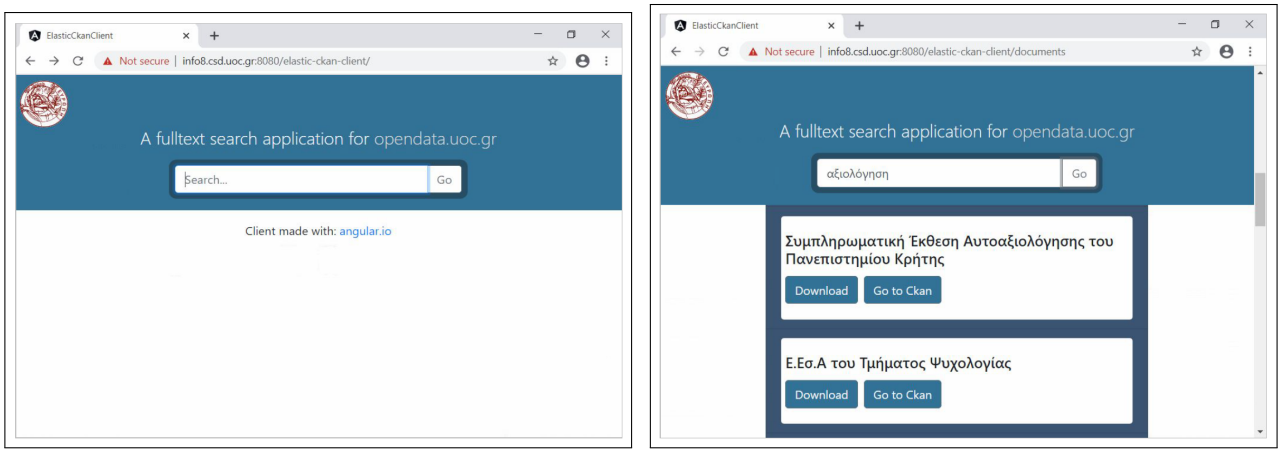

Figure 5 Full text search example results

Table 4 UoC student population historical datasets

Name of the dataset	Number of records
Graduatestudents_academicyear	3,275
Graduatestudents_graduationyear	2,608
Phds_academicyear	6,003
Phds_graduationyear	4,845
Undergraduatestudents_academicyear	8,030
Undergraduatestudents_age	12,043
Undergraduatestudents_graduationyear	6,746

All the above data are available for every Academic Department of the University of Crete (16 Departments) and from 2006 until 2019 (13 years). An indication of the size of each dataset is shown in Table 4.

After processing the original XLSX files, all data are stored in a MySQL database. The administrator can load new or update existing datasets to the DB from a web interface at any time. The REST API was implemented using Jersey[39], which is an open source Java framework for developing RESTful Web services. A visual documentation of the REST API was also implemented using the open source framework Swagger[40], which mainly helps developers and simplifies the consumption of the API by the applications (client side) and the implementation of the back-end. With Swagger we can easily design the API, test it and automatically generate, visualize and maintain the API documentation, ensuring that it stays up-to-date as the API evolves. Some example API operations regarding PhD students are shown in Figure 6

The Swagger UI allows the user to execute GET requests and experiment with different input parameters. As a response to an API call, the data can be returned in any of the following four output formats XML, JSON, (X)HTML and CSV, e.g. see Figure 7.

In addition, we implemented a web-based client application that meets this resource-oriented architecture, directly utilizes the developed REST APIs and provides search, filtering and data visualization services for the student population historical data (see Figure 8). The client application was also developed in Java using the Jersey framework. Any visitor is able to browse the available data in a simple and easy way, and the returning results can be visualized either in the form of tables (Figure 9a) or in the form of graphs (Figure 9b). For the dynamic creation of the graphs, the `D3.js`[41] Javascript library was used. More specifically, the requested data are generated in CSV format, which can be managed and displayed by `D3.js`

6.3 *Bulk Import and Export Services and Connections with External Systems*

Currently we are working on the communication and integration with `http://data.gov.gr`, which is the central directory of public data of all Greek government agencies and contains more than 10K datasets and 340 organizations, as well as connection with other repositories of scientific data.

Instead of directly loading datasets to that catalog, an approach that would require manual effort and may suffer from inconsistencies, we investigate an alternative approach, where our catalog remains to be the central management unit, and auxiliary tools are used for selecting what datasets from our catalog to publish to the central national catalog and for performing the necessary transformations.

To this end, we have developped a Java web application that utilizes the programmatic interface of our Open

[39] `https://eclipse-ee4j.github.io/jersey/`
[40] `https://swagger.io/`
[41] `https://d3js.org`

Data Catalog and data.gov.gr (they both use CKAN as the underlying platform) and supports the following:

(a) Automatically exports selected datasets from our catalog, that comply with specific matching criteria, and directly imports them to the http://data.gov.gr. For these purposes, different filtering options are available, e.g. category of dataset, department of origin, specific tags, date range etc.

(b) Makes the required transformations and renaming to the exported datasets so that they can be clearly and fully labeled when imported to data.gov.gr and therefore easily identified and searchable (for example, including the organization name into the dataset name or into the metadata).

(c) Logging and provenance of the whole update/export/import procedure.

We are also considering to provide a bulk and/or automated dataset upload (and export) procedure, combined with a custom data interface to other internal UoC systems, e.g. the University Student Registry (anonymized data), the Research Directory and other administration and monitoring subsystems that generate useful and open datasets. All of the above could provide a level of integration that would eliminate direct human intervention during the open dataset creation.

7 Possible Next Steps

Here we describe three main directions for next steps.

7.1 Production of Knowledge Graph and Linked Data

An interesting direction is the production of Knowledge Graph based on the contents of all datasets. Of course, this requires applying methods for large scale integration, see [11] for a recent survey. Having all data in the form of a knowledge graph will enable the users to browse all information in a coherent and semantically meaningful way. Moreover, users or other applications could issue SPARQL queries for getting tabular data and then perform various analyses or visualisations. Apart from providing browsing and query services, it would be sensible to provide also a user friendly search system over RDF data, e.g. the multi-perspective keyword search services described in [9]. In the ideal case, the knowledge graph could also contain information about the research output of the university, like [6]. That would enable the exploitation of these data for measuring various indicators, and thus alleviating the cost required for collecting data for computing these indicators (towards the direction of [4]). Currently such needs (which are periodic) are covered either by manual collection of data by each university, and/or by resorting to academic search engines, see [8]

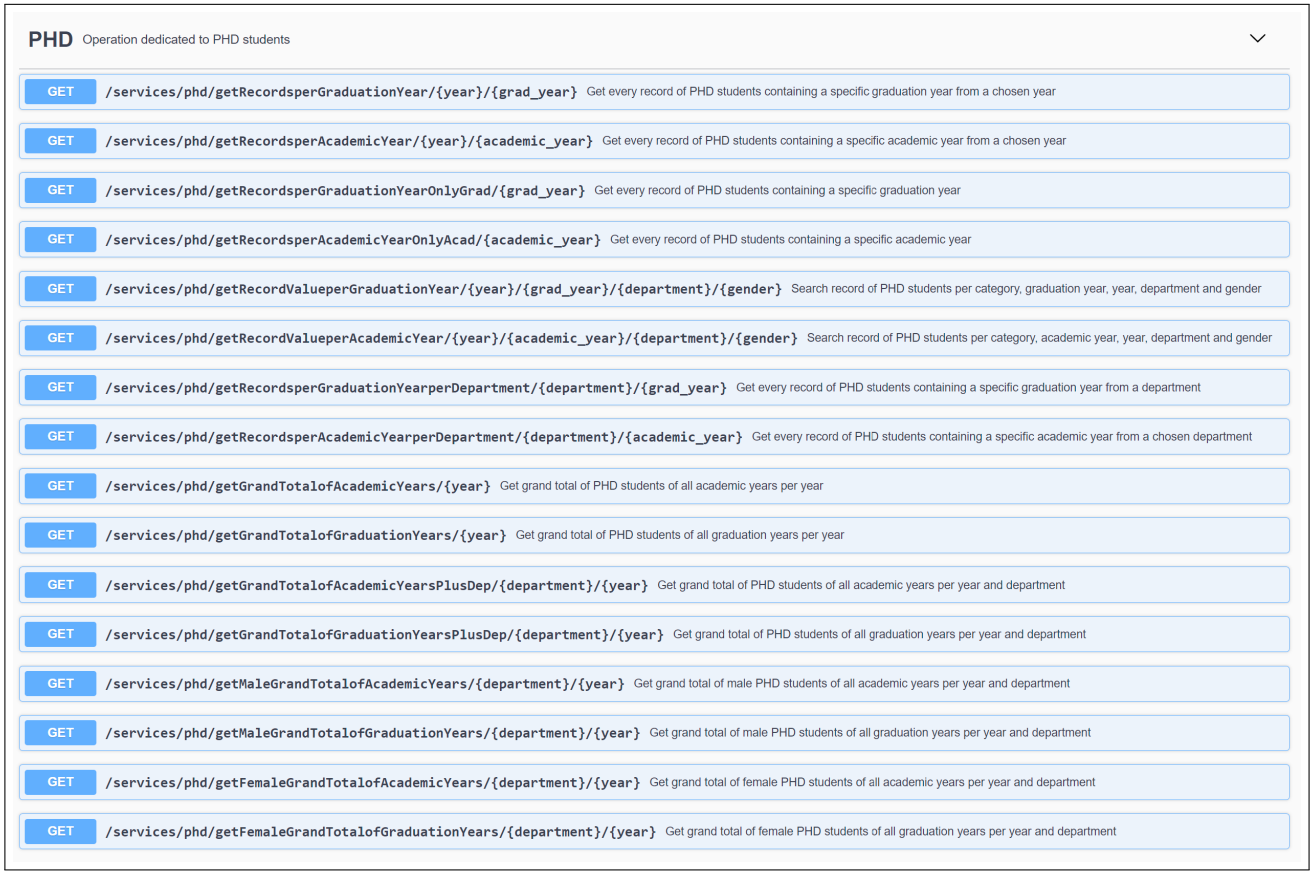

Figure 6 Available GET requests and their syntax regarding PhD students

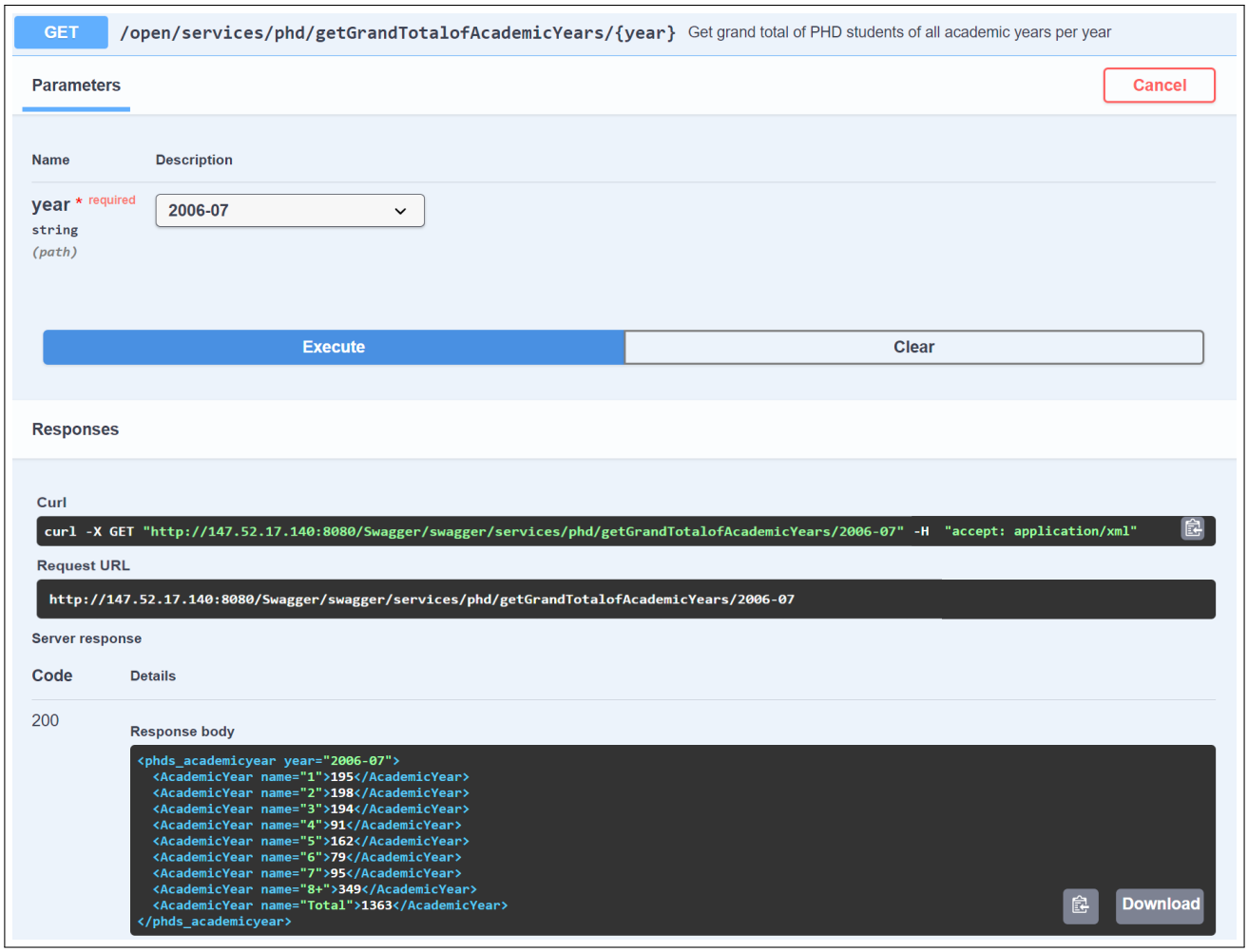

Figure 7 Example GET response in XML format

for a recent comparison of 12 academic search engines and bibliographic databases.

7.2 Repository of Scientific Data

As regards repositories for scientific data, HELIX (the Hellenic Data Service)[42] is the national e-Infrastructure in support of data-intensive research, handling the data management, analysis, sharing, and reuse needs of Greek scientists, researchers and innovators. HELIX provides its services directly to its users as an autonomous cloud infrastructure in support of data sharing, open access publishing, and data experimentation, in a horizontal manner to thematic and domain-specific infrastructures and communities. In future, and as this initiative progresses, we plan to investigate how to connect our catalog with that platform.

However a related question is "when are researchers willing to share their data" and for this question [15] presents the results of a broad quantitative analysis including around 1000 researchers from 13 large to medium-sized universities in Germany that discusses 11 hypotheses regarding researchers' intentions to share their data.

[42] https://hellenicdataservice.gr/

7.3 Training and Good Practices

Good practises for each individual type of dataset (news, personnel lists, etc), for increasing the percentage of datasets that are in a machine-processable format (recall the KPIs described in Section 5.3). This is also related to suggestions for increasing the interoperability of the information systems of the university, so that some information to be automatically exported in the catalog without any human intervention. Finally, it is important to provide incentives for providing complete, accurate and machine process-able data. Currently we are using the the list of datasets produced in Step $\mathcal{S}_{③}$, i.e. those datasets listed in Decision on Open Data that was formally approved by the University (April 18, 2019) and was submitted to the Ministry of Education and Religious Affairs (that for each academic department contains the datasets listed in Table 3), for measuring the completeness of the data in the websites of the Departments and units of the university.

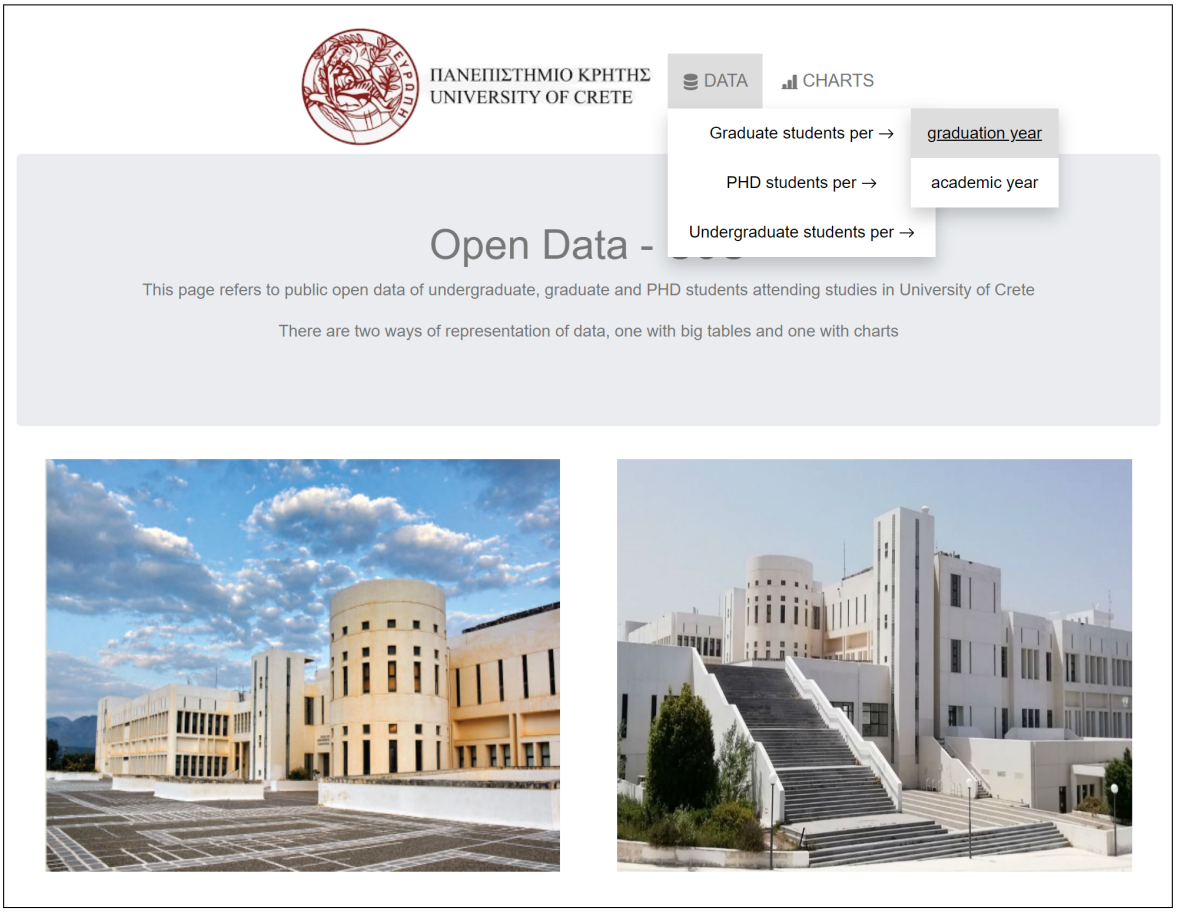

Figure 8 Home page of REST client web-based application

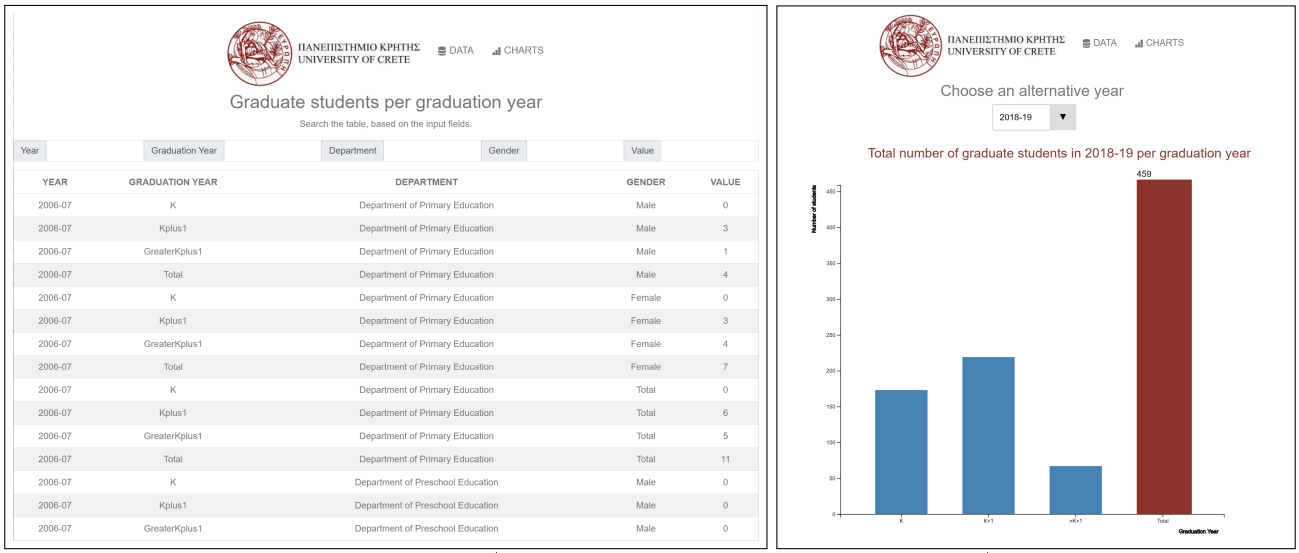

Figure 9 Left: Example tabular visualization (postgraduate students per year of graduation), Right: Example graph visualization (number of graduate students per graduation year)

8 Concluding Remarks

Even though organizations benefit from open data, the provision of an organizational and technical framework for Open Data in a big organization, like a university, is a challenging task. In this paper we described related work, worth-noting cases, and provided an analysis of the open data of Greek Universities. Then we described the five-steps process that the University of Crete has followed and the outcomes of this process so far, as well as the forthcoming steps. In brief, we detailed how the platform was selected, how we customized the platform, how we organized the catalog in department/units, what metadata we selected to include, what access services are offered, and an analysis of the uploaded content so far. A key point is that we adopted a mixed, bottom-up and top-down approach in which the catalog played a central role in the entire process. Even if the majority of the data are not in a common, easily processable form, we have realized that the catalog also serves as a global index of the various resources that are published in the various websites of the university. To set up priorities and plan our next actions, we have identified a few KPIs. We then described extra systems and services that we implemented that were required but are not covered by the catalog platform (content-based search, REST APIs to popular datasets that contain statistic information and import and export services). We hope that this to be useful in other organizations that have similar obligations and characteristics.

Acknowledgement

Many thanks to the undergraduate students who have contributed to this work, namely Skepetzis Giorgos, Koumourou Sotiris, Karkoulis Ioannis and Theofanous Panagiotis.

References

[1] Spiros Athanasiou, Vasilis Amiridis, Maria Gavrilidou, Evangelos Gerasopoulos, Alexandros Dimopoulos, Georgina Kaklamani, Fotis Karagiannis, Iraklis Klampanos, Dimitra Kondili, Kostas Koumantaros, Panos Konstantopoulos, Katerina Lenaki, Angelos Likiardopoulos, Natalia Manola, Despoina Mitropoulou, Agiatis Benardou, Nikolaos Boukos, Alexandros Nousias, Maria Ntaountaki, Stella Oikouta, Lia Ollandezou, Elli Papadopoulou, Ilias Papastamatiou, Stelios Piperidis, Orestis Speyer, Prodromos Tsiavos, Evangelia Chrysina, and Fotis Psomopoulos. National plan for open science. Technical report, January 2020. https://doi.org/10.5281/zenodo.3908953.

[2] Christine L Borgman. Open data, grey data, and stewardship: Universities at the privacy frontier. *Berkeley Tech. LJ*, 33:365, 2018.

[3] Tim Coughlan. The use of open data as a material for learning. *Educational Technology Research and Development*, 68(1):383–411, 2020.

[4] Cinzia Daraio and Andrea Bonaccorsi. Beyond university rankings? generating new indicators on universities by linking data in open platforms. *Journal of the Association for Information Science and Technology*, 68(2):508–529, 2017.

[5] Tobias Enders, Carina Benz, Ronny Schüritz, and Pamela Lujan. How to implement an open data strategy? analyzing organizational change processes to enable value creation by revealing data. In *ECIS*, 2020.

[6] Michael Färber. The microsoft academic knowledge graph: A linked data source with 8 billion triples of scholarly data. In *International Semantic Web Conference*, pages 113–129. Springer, 2019.

[7] Miriam Fernandez, Mathieu d'Aquin, and Enrico Motta. Linking data across universities: an integrated video lectures dataset. In *International Semantic Web Conference*, pages 49–64. Springer, 2011.

[8] Michael Gusenbauer. Google scholar to overshadow them all? comparing the sizes of 12 academic search engines and bibliographic databases. *Scientometrics*, 118(1):177–214, 2019.

[9] Giorgos Kadilierakis, Christos Nikas, Pavlos Fafalios, Panagiotis Papadakos, and Yannis Tzitzikas. Elas4RDF: Multi-perspective Triple-centered Keyword Search over RDF using Elasticsearch. In *(Demo Paper) Proceedings of the 17th Extended Semantic Web Conference (ESWC'2020)*, 2020.

[10] Rouzbeh Meymandpour and Joseph G Davis. Ranking universities using linked open data. In *LDOW*, 2013.

[11] Michalis Mountantonakis and Yannis Tzitzikas. Large-scale semantic integration of linked data: A survey. *ACM Computing Surveys (CSUR)*, 52(5):1–40, 2019.

[12] Markus Perkmann and Henri Schildt. Open data partnerships between firms and universities: The role of boundary organizations. *Research Policy*, 44(5):1133–1143, 2015.

[13] Sergi Rovira, Eloi Puertas, and Laura Igual. Data-driven system to predict academic grades and dropout. *PLoS one*, 12(2):e0171207, 2017.

[14] Erna Ruijer, Stephan Grimmelikhuijsen, Jochem Van Den Berg, and Albert Meijer. Open data work: understanding open data usage from a practice lens. *International Review of Administrative Sciences*, 86(1):3–19, 2020.

[15] Stefan Stieglitz, Konstantin Wilms, Milad Mirbabaie, Lennart Hofeditz, Bela Brenger, Ania López, and Stephanie Rehwald. When are researchers willing to share their data?–impacts of values and uncertainty on open data in academia. *Plos one*, 15(7):e0234172, 2020.

[16] Yannis Tzitzikas, Marios Pitikakis, Giorgos Giakoumis, Kalliopi Varouha, and Eleni Karkanaki. How can a university take its first steps in open data? In *14th International Conference on Metadata and Semantics Research, MTSR'2020*. Springer, 2021.

[17] Jose Zubcoff, Llorenç Vaquer Gregori, Jose-Norberto Mazón, Francisco Maciá Pérez, Irene Garrigós, Andrés Fuster-Guilló, José Vicente Cárcel Alcover, et al. The university as an open data ecosystem. 2016.