Open edX platform
Global footprint report
December 2024

v1.1

by Juan Camilo Montoya

Core contributor

This report has been prepared as a contribution to AXIM collaborative and the Open edX initiative.

1. Executive Summary

This report presents a comprehensive analysis of the global footprint of the Open edX platform. It has been prepared by Juan Camilo Montoya with the support of other members of the edunext team[1] as part of our core contributions to advancement of the Open edX project. This analytic exercise aimed to quantify and categorize Open edX deployments around the world, drawing on multiple data sources and employing both automated and manual data processing techniques. Major findings include the total number of sites/instances, geographic distribution, language breakdowns, types of initiatives deploying Open edX, and trends in adoption over time. These insights help illustrate the platform's reach and serve as a basis for future improvements in tracking and community engagement.

---

2. Introduction

The Open edX platform has been since 2013, and specially in the last few years positioning itself as one relevant technology for delivering massive open online courses (MOOCs) and high quality online learning experiences. As part of this process, understanding its global footprint is crucial for stakeholders within the project—including AXIM collaborative, Open edX providers, institutions using or planning to use the platform, and the broader Open edX community—to gauge platform adoption, identify usage trends, and uncover new opportunities for collaboration.

Edunext has undertaken this research to provide a snapshot of Open edX usage worldwide, offering actionable insights into how and where the platform is being utilized. This report details the methodology applied, the data sources referenced, the analyses performed, and the challenges and future steps that will guide continued improvements in tracking and reporting Open edX adoption.

---

3. Data Sources

Our analysis draws upon information from a variety of sources to ensure comprehensiveness and accuracy:

---

[1] Special acknowledgement to the work of Felipe Montoya, Daniela Ríos and Andres Espynel.

- **BuiltWith**: A technology profiler service that identifies platforms and frameworks used on websites. BuiltWith provided an initial list of Open edX-powered sites, as well as metadata on these sites' technology stacks.
- **EduNext Internal Data**: As a leading Open edX provider, edunext maintains records of client and partner implementations. This dataset offered detailed insights into hosting arrangements, deployment details, and usage patterns.
- **Historical Lists Maintained by Axim**: These lists contained older, partially verified, yet useful information regarding Open edX instances identified over the years. Although some entries may be outdated or missing details, they contributed a good source for data aggregation prior verification.

---

4. Methodology

4.1. Sanitization of Information

1. **Aggregation of Sources**: We compiled data from BuiltWith as of november 30th 2024, edunext's internal records as of December 31st 2024, and the available Axim provided list into a single master dataset.
2. **Deconcatenation**: In the specific case of Buildwith, which provides a very large number of records, these are aggregated by parent domain, which may be misleading in some cases where multiple actually different instances share the parent domain, so we created an automated process for splitting these aggregated records into separate records for higher accuracy.
3. **Deduplication**: Entries from multiple sources often overlapped. We implemented an automated deduplication process to remove duplicates.
4. **Removal of sites outside of the scope:** a number of specially crafted criteria was applied to the master dataset in order to remove all the domains that were outside of the sphere of interest for this analysis, for example testing sites or prototypes in the edunext multitenant solution in the free tier, studio and preview domains when the lms domain was already on the list, staging environment domains, etc.

4.2. Automated domain validation via Web Scraping

We performed a programmatic web scraping technique on each domain on the sanitized list to verify the 2 main qualification criteria for this study:

of each unique domain to confirm key attributes:

- **is_online**: Checking whether the domain is responsive at the time of scraping.
- **is_openedx**: Verification that the site is actually running the Open edX platform based on the validation of the response to the “/heartbeat” endpoint.

One of the challenges faced in this stage is the fact that the open edX platform can sometimes be hidden behind layers of marketing sites or authentication paywall, which can exclude from the analysis certain initiaitves.

4.3. Automated platform specific datapoints collection via Web Scraping

With the resulting validated dataset, we performed a programmatic web scraping technique that focused on Open edX specific features:

- **Initiative Name:** Inferring the way the platform is configured to refer to itself.
- **Open edX Release**: Determining the specific version of Open edX that is running.
- **num_courses**: Inferring the number of courses is visible in the catalog, if such endpoint is available.
- **Number of courses per year**: Where feasible, establishing timeframes of site operation (Number of courses starting each year.)
- **Language:** The language in which the page is served.
- **IP**: Collecting IP addresses of the webservers.

4.4. Data Inference of Initiative specific datapoints

Leveraging the data this far, and the advanced reasoning and web search capabilities of frontier LLMs models, we engineered a RAG+search automated classification method to make additional inferences in some key aspects of the initiative, including:

- **Initiative Type**: Categorizing the implementing organization as a university, corporate, government body, NGO, etc.
- **Initiative Size**: Estimating the size or scale of the initiative based on public information, including user enrollments or institutional descriptions.
- **Geographic location**: Inferring geographic location at the country level domain specific data and textual cues from site content.
- **Operational Status**: Classifying the site’s operational status (active, archived, under maintenance, etc.).

4.4. Quality Control

Our team used a combination of automated scripts and manual reviews to validate the data and spot any potential error. This included:

- Cross-verification of records with public registries.
- Spot-checking the results from the ai based inference for accuracy.
- Reviewing older data for consistency with current records.

4.5. Analysis

Once the data was consolidated and validated, we performed statistical and exploratory analyses to identify patterns and trends for the different dimensions included in this analysis. The results are detailed in the next section.

---

5. Results of the Analysis

5.1. Total Number of Sites and instances

This exercise allowed us to identify 2300 different and confirmed Open edX Sites.

It is key to note that the Open edX platform does support multitenancy features, meaning that multiple sites can be hosted in the same instance.

When extracting only the number of different instances as identified with a different IP address, the number of Open edX instances rises to 1266.

5.2. Analysis of the Main Multisite Instances

A small number (12 instances) were found to have more than 10 sites sharing the same IP address or pool. The most prominent examples are:

- edunext multitenant instance: 468 sites
- lms.futurex.sa: 120 sites
- Moocit multitenant instance: 46 sites
- Project-eu : 31 sites

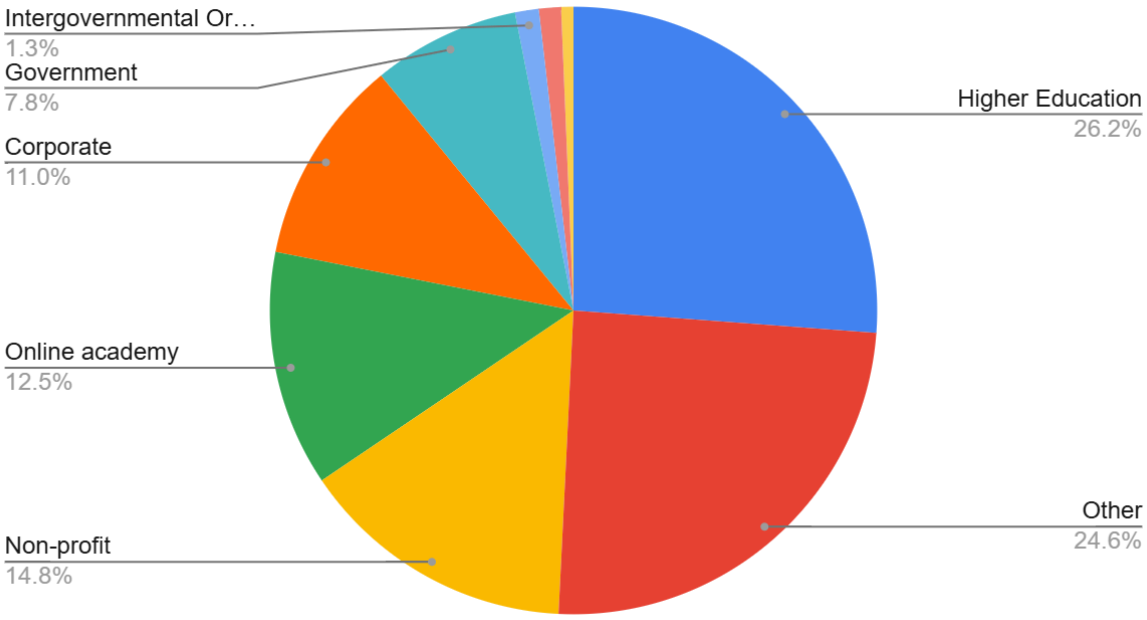
This methodology did not pick up on some well known multitenant instances such as appsembler, edly.io, edspirit. More research and collaboration with those actors will be needed to complement and improve this exercise.

5.3. Breakdown by Inferred Initiative Type

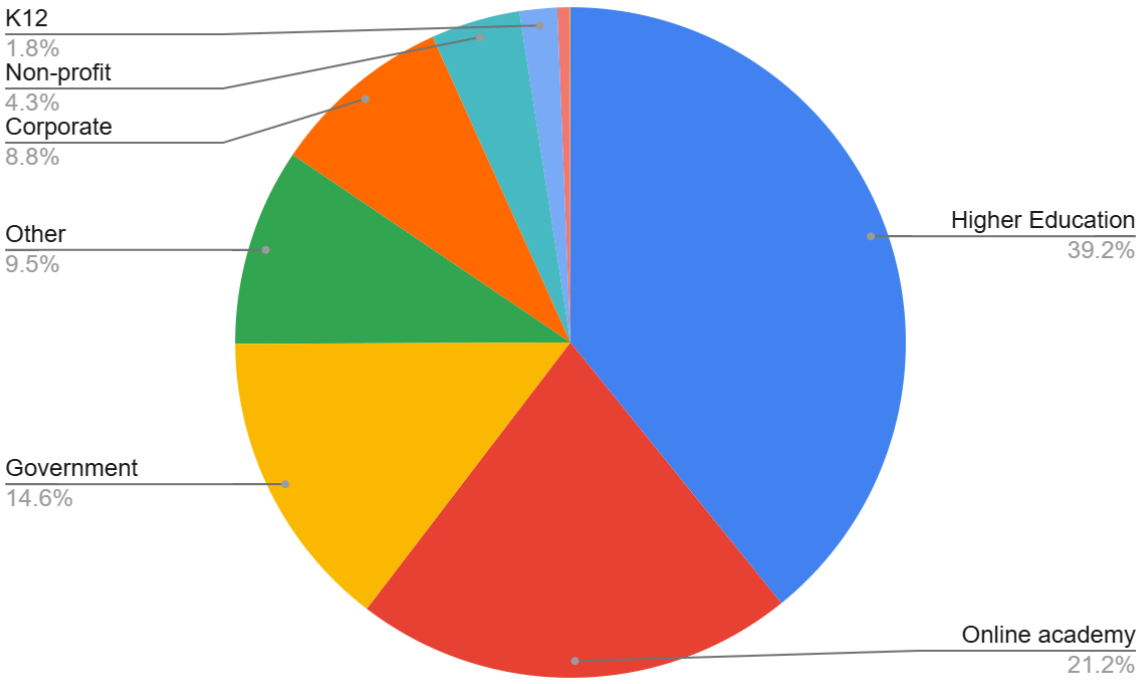
When breaking down the initiatives by their inferred type, the most frequent categories are Higher education, non profit, online academies and Corporate initiatives.

An effort was made to differentiate more granularly this category, using the criteria below:

- **K12** – Geared toward children from kindergarten through high school
- **Higher Education** – Targeting adult learners, developed or sponsored by higher education institutions
- **Government** – Part of a government program or agency
- **Intergovernmental Organization** – Part of a multilateral organization or agency
- **Online academy** – Offering public training, possibly paid or monetized
- **Corporate** – Driven by a company for employee/customer training
- **Non-profit** – Driven by a non-profit, foundation, or charity (often free courses)
- **Religious** – Focused on religion or spiritual content
- **Other** – Does not clearly fit any category above



If the total number of course runs is considered instead of the number of sites, the share of the Higher education type sees an increase to 39.2% and the share of Government initiatives rises to 14.6%, as shown in the graph below.



5.4. Breakdown by Geography

This is the breakdown of the number of sites by region:

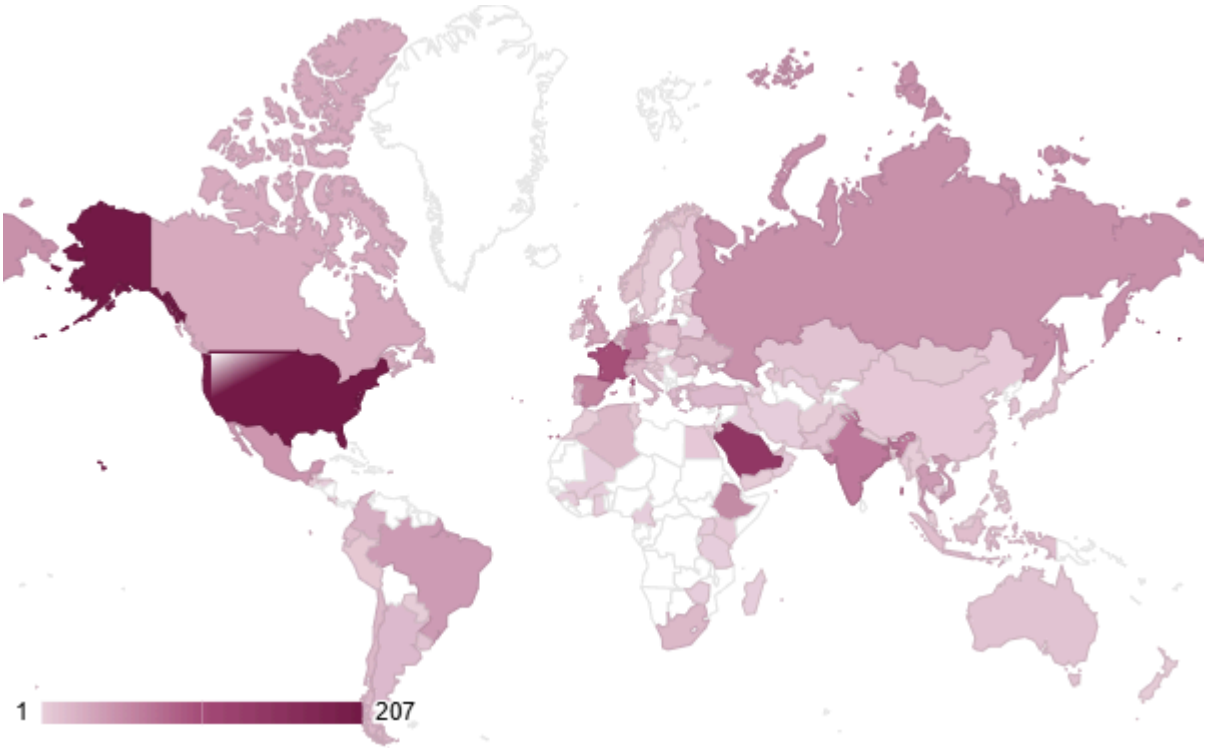
Region	number of sites	average number of course runs
unknown	733	69.8
Western Europe	436	45.3
North America	280	242.1
Mena	207	214.9
East Asia	182	129.6
South America	132	44.3
Eastern Europe	114	72.5
Africa	90	134.8
South Asia	86	279.2
Oceania	12	45.0
Central America & Caribbean	11	23.2
Central Asia	10	45.6

While the most number of sites can be traced back to countries in western Europe, they have an average of only 45.3 course runs. In North America instead, the number of sites is smaller but the average number of courses is

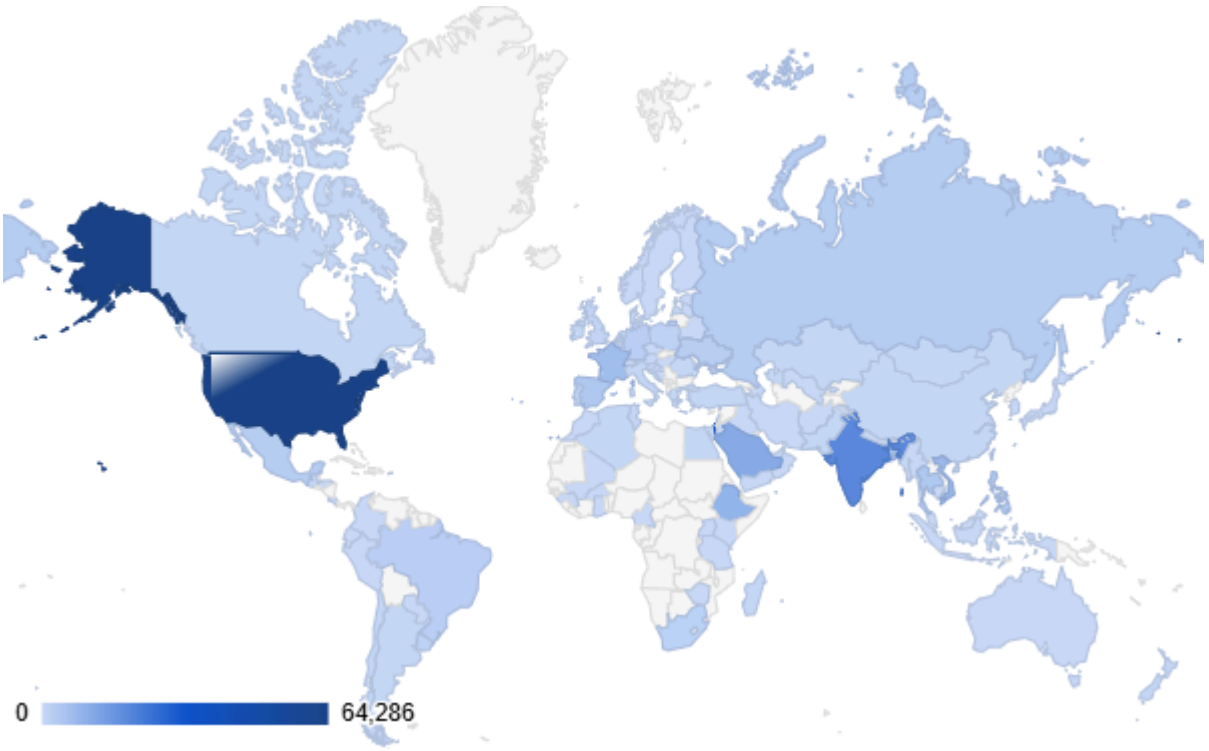
much larger (242.1). Even if bit outliers as courses.edx.org and edge.edx.org are not included, the north america region still has a high average of 105.3 course runs per site.

The region with the largest average number of course runs is South Asia.

In terms of distribution by country of the number of sites, here is the geographic representation of the Open edX footprint:

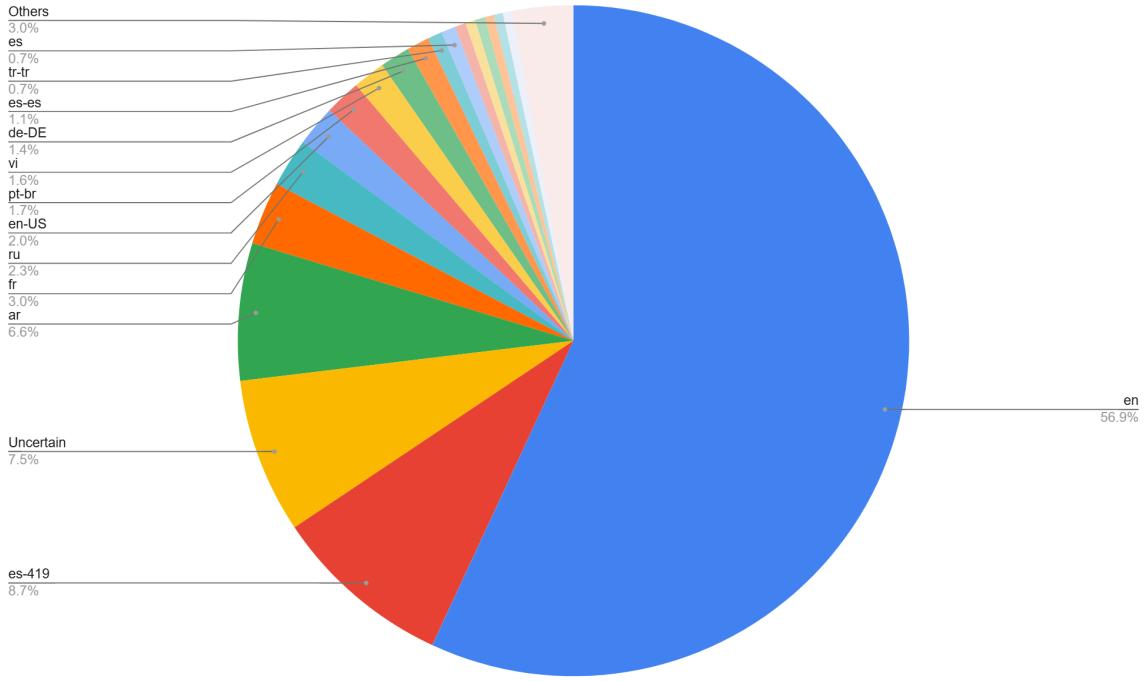


When viewed by number of course runs, only minor differences can be seen in the distribution:



5.5. Breakdown by Language

The top 3 most prevalent languages are English, Spanish and Arabic. Here is the complete breakdown of the different languages used by the sites:



5.6. Breakdown by Inferred Size

This is one of the hardest inferences to make, given the fact that the learner population can only be proxied by indirect metrics such as the number of courses or the dimensions of the organization.

The numbers we landed with are presented in the table below:

Size	Number of sites	Average number of course runs
Very Large – More than 10,000 learners	308	569.0
Large - 1,000 to 10,000 learners	125	122.3
Medium – 100 to 1,000 learners	692	31.0
Small – 50 to 100 learners	128	6.1
Very Small – Fewer than 50 learners	234	1.7
Unknown	903	55.9

However, much better tooling or manual verification would be needed in order to provide better estimates in this category.

5.7. Total number of course runs

The estimation of the number of course runs in each site for this exercise was done based on the number of course runs that are configured to be visible by the course api endpoint. This is an imperfect metric due to the following reasons:

- It may hide course runs that are fully operational, but for some reason are not configured to be visible.
- When using an additional marketing site to handle the catalog of courses, the platform administrators may not attend to the configuration of the course API visibility status for many course runs, and thus the endpoint will display courses that are not really part of the initiative's offering.

Nevertheless, the data we can collect allows some initial interpretations.

The aggregated number of course runs reported from this dataset of sites adds up to 257.688 course runs.

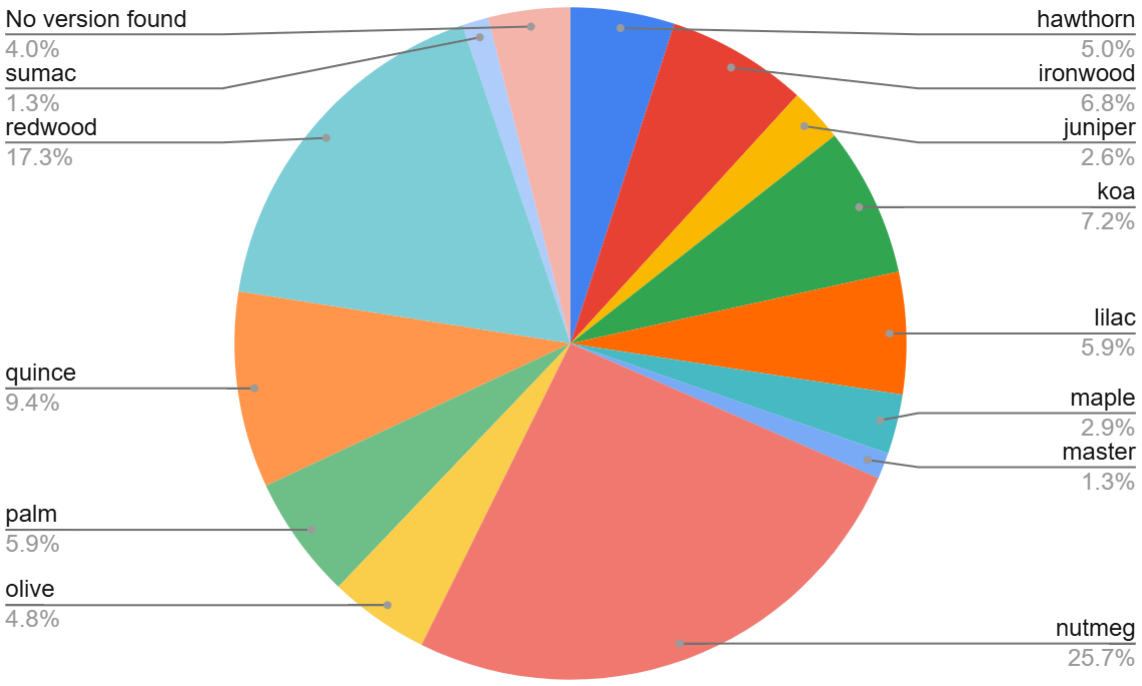
More than half of these course runs are found in the top 20 records in the list by number of course runs:

openedx_domain	name	SUM de num courses

courses.campus.gov.il	Campus IL	31422
courses.edx.org	edX	25800
edge.edx.org	edX	12555
mutaaheb.srca.org.sa	متأهب	9502
lms.ricesmart.in	RiceSmart	6718
lms.hutech.edu.vn	HUTECH eLearning	6425
latam-myacademy.learning-tribes.com	My Academy learning tribes	4964
apac-myacademy.learning-tribes.com	My Academy learning tribes	4871
bux.bracu.ac.bd	buX | BRAC University	4098
emea-myacademy.learning-tribes.com	My Academy	3943
credocourseware.com	Credo Learning Tools	3915
all-courses-pa.pearson.com	Pearson	3796
learning.rpsconsulting.in	RPS Academy	3782
certprepcourseware.pearson.com	Pearson CertPREP	3145
lms.fun-mooc.fr	FUN France Université Numérique	3130
lms.nimblywise.com	Weave	2245
courses-api.openedu.tw	首頁 | 中華開放教育平台	2043
lms.ust.edu	UST	2033
class.eduquestph.com	EduQUEST	2000

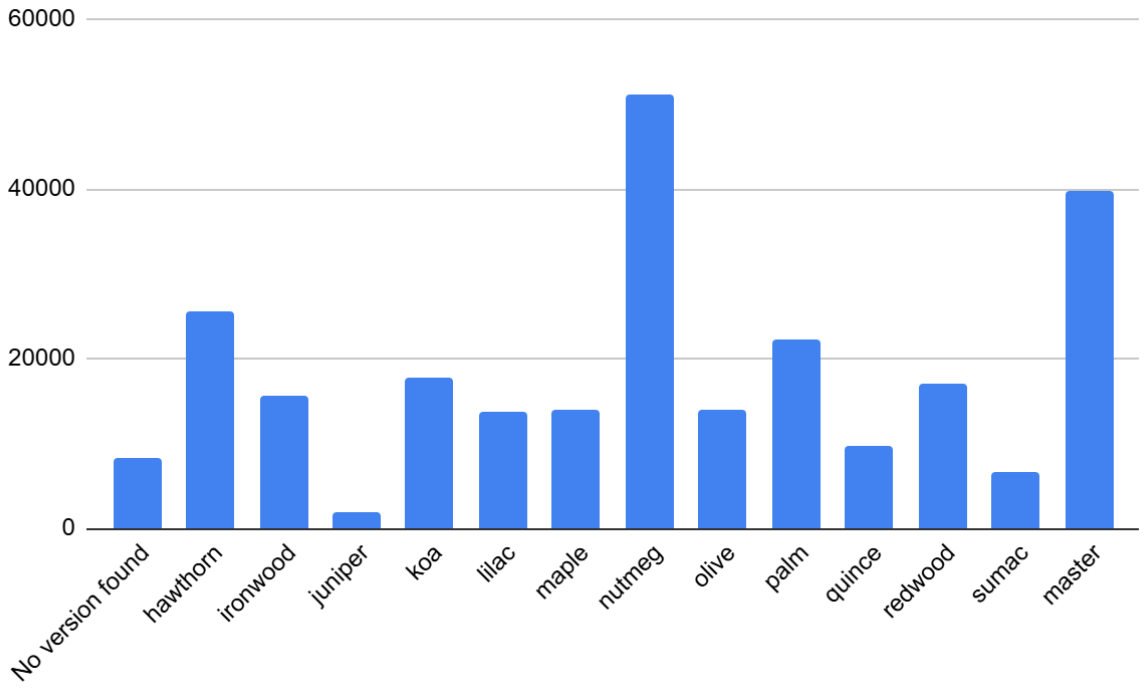
5.3. Breakdown by Open edX Version

When breaking down the sites by their reported Open edX version, you can see that despite the past 3 releases already gaining significant adoption, a very large number of sites are still hosted in releases from Hawthorn to Nutmeg.



This supports the idea that migrating to new versions of the platform is for many initiatives very hard or costly, and specially when migrating from version up to Nutmeg where key paradigm shifts such as MFEs and k8s are not yet implemented.

When considering the Open edX releases by the number of course runs, this is the distribution:



Special considerations about this numbers:

- edunext multisites instance, representing 468 sites, currently runs in Nutmeg and Campus IL, the largest instance by number of course runs is in Nutmeg.
- large initiatives such as futureX and the e-SHE that collectively represent approximately 200 sites run in Redwood
- Moocit, representing 46 sites runs in Ironwood
- Project eu, representing 31 sites runs in Hawthorn
- 4 of the 20 largest initiatives by number of courses run in Palm (lms.ricesmart.in, credocourseware.com, apus.credocourseware.com, lms.nimblywise.com)
- Although there are 29 sites reportedly using the “master” release, 96% of the course runs in this group belong only to 2 sites, courses.edx.org and edge.edx.org

5.9. Breakdown by Inferred Operational Status

This is also a challenging inference, but given that there is a significant number of Open edX sites that exist in suboptimal state, we’ve made an effort to differentiate them by this categorization, using the criteria below:

- **Prototype or Test Site** – the site has minimal content, it explicitly says it is under construction, or is labeled demo/sandbox/staging
- **Operational with Issues** – it has Some course offerings but also has notable problems (broken links or images, outdated info)
- **Operational** – Functional, actively offering courses with updated content

And these are the resulting metrics:

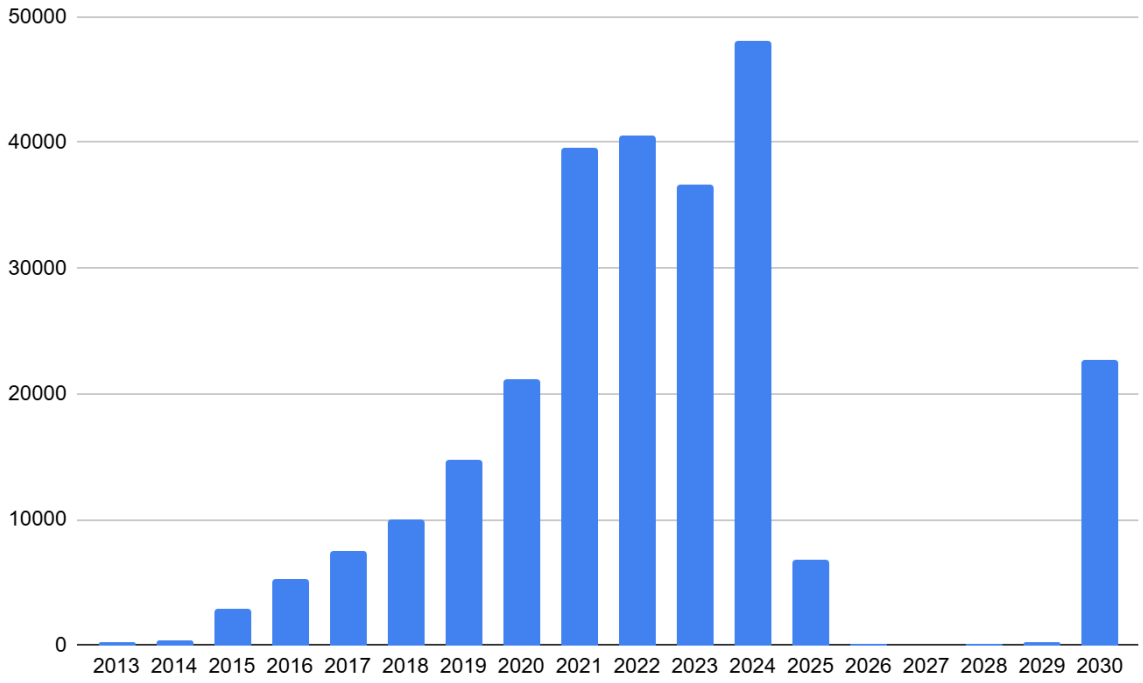
Operational status	Number of sites	Average number of course runs
Operational	1449	158.7
Operational with Issues	269	62.8
Prototype or Test Site	561	9.1
Unable to determine	21	403.0

5.10. Analysis of Time Evolution

In order to estimate the general trend of usage of the Open edX platform over the year, we propose an alternative based on the information available in the /courses api.

As it turns out, this api endpoint delivers all the list of course runs in the catalog with certain metadata for each course run, and this metadata includes the date

when the course starts. By tracking and aggregating the number of course runs that are set to start in one particular year, we were able to infer a trend of platform usage across all sites as follows:



Note that the spike in the year 2030 is consistent with the fact that this is the default start year for newly created courses.

---

6. Challenges and Future Work

While this analysis provides a valuable snapshot of Open edX adoption worldwide, several challenges and areas for improvement remain:

- **Aggregation of Information from Community Sources**: Relying on external databases and community-maintained lists introduces the risk of incomplete or inconsistent data. Enhanced collaboration with the Open edX community and providers could improve coverage and accuracy.
- **Better Collection of Stats and Vitals from Application Endpoints**: Automated methods to query sites for usage metrics or performance indicators could lead to more precise data, specially on course offerings, enrollments, course completions, registered users, active users, etc.
- **Continual Updating of Data**: Given the dynamic nature of online platforms, frequent re-scraping and verification will be essential to maintain an up-to-date view of the global Open edX landscape.

- **Improvement of Inference Techniques**: While LLMs proved useful, refining the engineering around this mechanism and validating inferred fields could further enhance reliability.

---

7. Conclusion

In summary, this analysis shows the broad and growing reach of the Open edX platform, spanning diverse regions, languages, and institution types. By consolidating multiple data sources and applying robust cleaning and inference methods, we provide a detailed panorama of global deployments. Going forward, continued collaboration with community stakeholders, combined with ongoing data enhancements, will help maintain a dynamic, accurate view of the platform's footprint. This information will ultimately support strategic decision-making, targeted community engagement, and further innovation in online learning.

---

8. Appendices

- **Appendix A**: Detailed master data list with all the inferences and including the confidence level estimated for the LLM based inferences as well as the explanations and sources used.