CumInCAD 2.0: A Redesigned Scalable Cloud Deployment

Towards higher impact with openness and novel features

Tomo Cerovsek[1], Bob Martens[2]
[1]University of Ljubljana [2]TU Wien
[1]tomo.cerovsek@fgg.uni-lj.si [2]b.martens@tuwien.ac.at

CumInCAD is a cumulative index of publications related to 'Computer Aided Architectural Design' (CAAD). It includes bibliographic data of approximately 12K records, which were predominantly derived from CAAD-related conferences, such as ACADIA, ASCAAD, CAADRIA, eCAADe, SiGraDi and CAAD futures. A brief historical overview of almost two decades of collaboration between the University of Ljubljana and the above-mentioned CAAD-associations is provided. After years of successful operation the previous interface became gradually outdated, which called for new developments to assure continuous support to open access to scientific knowledge. In this contribution, we explain the existing status of the systems, its use, and the transition process to a cloud deployment.

Keywords: *Open access, Cloud deployment, Bibliometrics, Google Scholar*

INTRODUCTION

The main goal that propelled the 'Cumulative Index of Computer Aided Architectural Design' (CumInCAD) was: (1) to build a 'collective memory' of scientific publications from conference proceedings of CAAD-associations; and (2) to make this memory unconditionally accessible to the scientific community as a web-based bibliographic repository of works.

The need for a digital repository of CAAD-related publications was initiated by two important issues:

- Possible data loss
- Limited access

Possible data loss. There was a fear that valuable scientific contributions would be lost or at least researchers' time could be wasted on unnecessary rework, or time-consuming literature research and information retrieval.

Limited access. Access to CAAD scientific works was very limited due to a small number (up to 250) of printed copies of the proceedings. Conference proceedings are seldom stored in libraries as complete series, but remain mostly in the participants' bookshelves.

Almost two decades ago, we responded to the need - to prevent data loss and to improve access to the works - and set up a web-based bibliographical repository, which used Perl as its programming language and a Web Oriented Database (WODA). This made bibliographic records and full texts available to a wide spectrum of interested researchers and practitioners. A positive response from the scientific community contributed to a growing reputation of the CumInCAD. As the reputation of CumInCAD grew, so did its content, encompassing today over 12,000 bibliographic records and more than 9,000 full texts.

Three important maintenance aspects for the status of CumInCAD and its survival are:

Maintenance of bibliographic data. The driving force for a continuing support for a commitment to collect, organize and make the content available via open access were volunteering activities by participating CAAD-associations and individuals. For quite some time, conference organizers have been using digital master files in pdf, which are later used for the repository along with bibliographic data.

- Past status: A time consuming preparation of bibliographic data with limited impact.
- Goal: To ease bibliographic data preparation with a helpdesk and to improve the impact.

Maintenance of the technical solution. Almost 20 years is a considerable amount of time for any IT system, especially in view of current developments in web-based (electronic) publishing. A growing user-base and amount of content on the one hand and continuous technological changes and advancements for the www on the other represent constant technical challenges for CumInCAD.

- Past status: The hardware and software system resided on a University of Ljubljana server that wasn't dedicated to CumInCAD only; e.g., some services may require an update, which could (and did) cause problems to CumInCAD.
- Goal: To migrate to a solution that is dedicated and not dependent on the institutional IT, but open, independent, and scalable.

Maintenance of the open access path. Over the last few years, we established a culture of 'Limited Open Access'. This means that CAAD-associations created tangible "added value" by making their recent full texts exclusively available to their members (read as registered CumInCAD users). The intention was not to generate business; on the contrary, the main aim was to keep the repository on a shoestring budget.

- Past status: Limited Open Access
- Goal: Full open access to CumInCAD.

METHOD: TRANSITION STEPS

As a response to the identified problems of the past status, the following line of action was set out:

Agreement on continuous support. We agreed to extend the collaboration between the University of Ljubljana and CAAD-associations to assure continued development, support and migration of the existing bibliographic repository. A migration plan was set up for the second half of 2015. A preliminary webometric and bibliometric analysis was performed to assess relevance and impact.

Set-up of a helpdesk system. In order to support both maintenance of the bibliographic data and migration, we set up a help desk that provides easy and professional reporting and tracking. The helpdesk system will also enhance end-user experience in the future and will remain an integral part of CumInCAD.

From limited to full access. CAAD-associations decided to upgrade the previous approach of Limited Open Access and to progress towards 100% Open Access. Open Access to everybody leads to enhanced visibility. In order to increase impact, we also decided to technically support google scholar crawling, indexing and search engine optimization.

Porting the system to the cloud. In order to avoid the problems related to the closed technical system that is affected by the hardware and software limitations and lack of portability, we decided to go for a dedicated, standalone, flexible cloud solution that will allow migration and scalable deployment.

Deployment, testing and re-launching. This part of the project encompassed a complete technical migration of the entire database content, services that were based on a common gateway interface, interpreters and supporting libraries, including tens of GBs of data. Before a final re-launch we also did a complete graphical re-design of the entire interface and performed tests on different configurations

In the next two sections of the paper we provide detailed information on the analysis and results, followed by conclusions and the plans for the way forward.

ANALYSIS OF USE PATTERNS

The analysis of use patterns is important as it helps us plan technical and organizational improvements of the web repository. For this purpose we used Google Analytics, a free web analytics service that provides statistics and basic analytical tools for search engine optimization (SEO) and marketing purposes. The service is available to anyone with a Google account.

Once the Google Analytics is active, it starts to collect, analyse and format the data to different visual representations. The dashboard data on Figure 1 give important information on the nature of CumInCAD end-users: we have over 76,000 users per year, 20% or approximately 15,000 returning users (i.e. they returned to the CumInCAD web site at least once). The y-axis shows page views (number of pages that were viewed per week) in the period of one year, starting on 1st of May 2015. One significant low-use period is during Christmas, and at the end of the time-scale (April 2016), where we see that the number of page views is approaching zero, because we completely ported the system to a new cloud solution at that time. Users usually view about three pages per session, which means that they find what they are looking for very quickly, or they quickly find out that the content is not appropriate for them. End-users usually stay on the site less than 2 min. We identified about 260 different sessions per day undertaken by approximately 200 different users.

As visible on a diagram we have one significant peak (high-use) of CumInCAD in September. As the number of users is not proportional to the page-views, we may assume that one of the end-users was a bot.

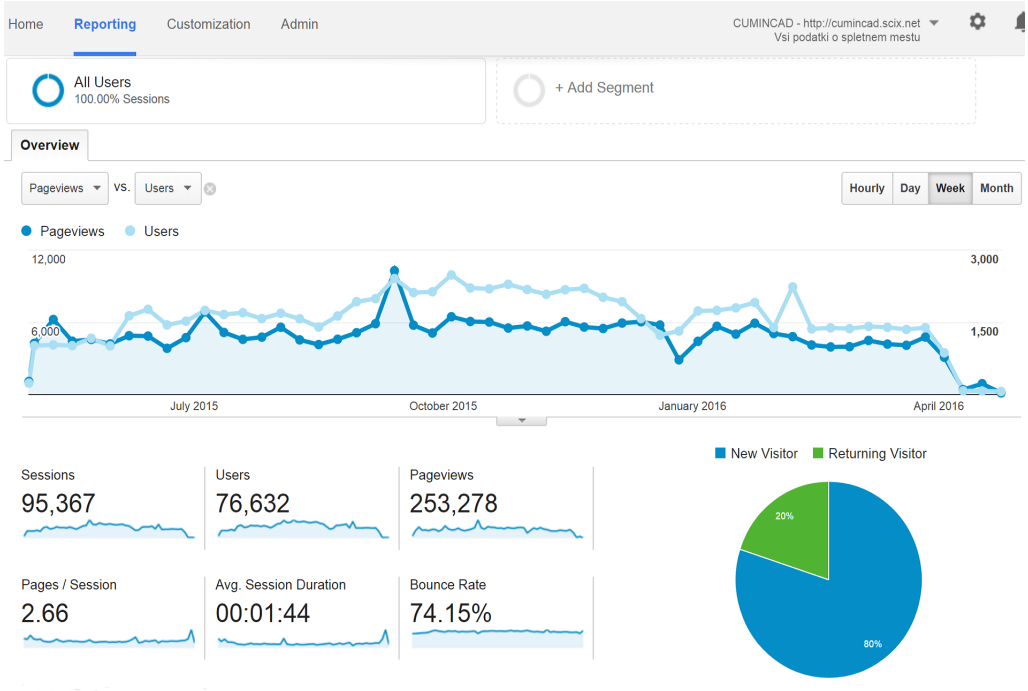
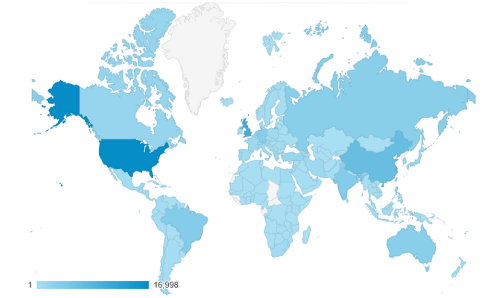


Figure 1
An overview of web access analysis for CumInCAD from 1. May 2015 to 30. April 2016



Primary Dimension: Country City Continent Sub Continent

Secondary dimension | advanced

	Country	Sessions	Sessions
		95,367 % of Total: 100.00% (95,367)	95,367 % of Total: 100.00% (95,367)
1.	United States	16,998	17.82%
2.	United Kingdom	10,073	10.56%
3.	China	6,879	7.21%
4.	India	3,950	4.14%
5.	Brazil	3,901	4.09%
6.	Germany	3,598	3.77%
7.	Australia	3,398	3.56%
8.	Russia	3,323	3.48%
9.	Canada	1,996	2.09%
10.	Austria	1,907	2.00%
11.	Netherlands	1,835	1.92%
12.	Philippines	1,828	1.92%
13.	Turkey	1,819	1.91%
14.	Japan	1,565	1.64%
15.	Taiwan	1,332	1.40%
16.	France	1,303	1.37%
17.	Italy	1,282	1.34%
18.	Iran	1,192	1.25%

Figure 2
Demographic analysis of web access by country from 1 May 2015 to 30 April 2016

As illustrated in Figure 2, almost 18% of the traffic comes from US; in terms of use of the system per capita, Austrians would be the most active end-users. We can also conclude that CumInCAD is truly international.

The referral analysis also shows that the end-users did not come to CumInCAD from Google, but they use direct referrals. This means that the implementation of Google Scholar would make an impact.

Based on the use patterns we may also predict how scalable a technical solution should be. On the other hand, we can use the Web analytics as an audit trail of the activities on the web systems and allows for the measure the affect the developments, and this is what we did during the migration and final transition to CumInCAD 2.0.

RESULTS OF MIGRATION

There are two important parts to migration: the work either invisible to the end-users (back-end) or visible in other systems (e.g., Google), and the work that is visible to the end-users as an interface (front-end).

Figure 3 presents an example of an effort that is not visible to the end-users of the CumInCAD directly, but through a Google Scholar interface (once CumInCAD was opened and ready for indexing, it was crawled and indexed by Google robots).

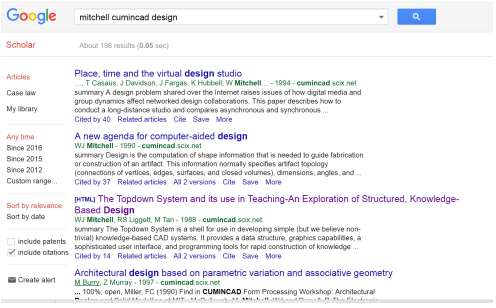


Figure 3
Google Scholar search results containing CumInCAD results

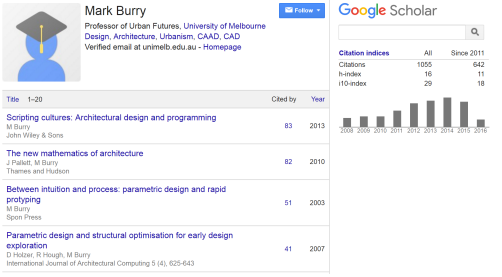
Back-end: Towards cloud deployment

We executed the following development steps towards open cloud deployment:

- Open access search engine friendly solution;
- Standalone portable system implementation;
- Cloud deployment of the system with backup.

Open access search engine friendly solution. Preparation of a pre-porting review, including bibliographic measures and review of web access stats and impact. These steps included adjustments of metadata. The indexing of the bibliography required adjustments of the metadata system, in addition we had to prepare dedicated index files for web crawlers that connect bibliographic records. Testing of the indexing on a Lucene Standalone Engine and on Google Scholar; implementation of a novel data model for indexing; preparation of initial indexing for the whole repository using pre-processing and XML records.

Figure 4
Google Scholar public display of author record with the display of bibliometric statistics



Search Engine Optimization (SEO) that includes indexing and referencing by way of the standard Google search engine to assure increased ranking. A trial implementation of the customized search approaches triggered technical decisions. We concluded this phase of migration by sending an invitation to CumInCAD users to create a Google Scholar account.

Google Scholar Citations provide a simple way for authors to keep track of citations to their articles. An author can trace back information about who is citing his/her publications, graph citations over time and compute several citation metrics. Authors can set up a public profile to appear in Google Scholar results when people search for them (see Figure 4).

Standalone portable system implementation. The second part of the transition was far more demanding as it required considerable technical modifications to be executed on time. This included porting of a database system to a new standalone. The database management environment used for input and output (data entry and automatic index file preparation) was adapted, tested and implemented.

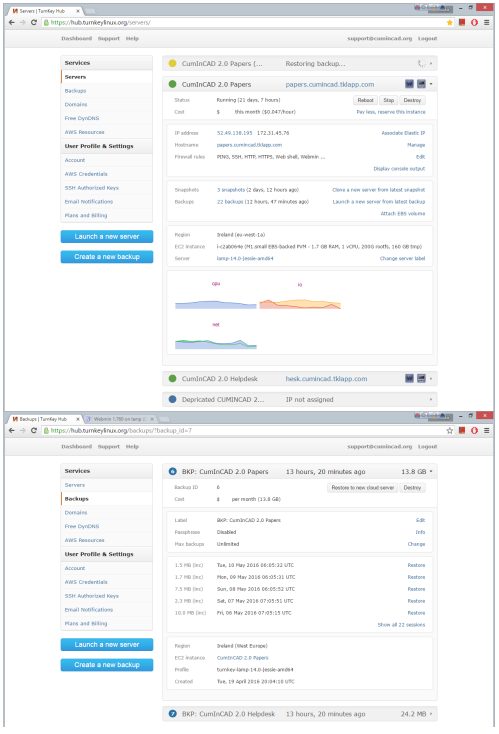
Cloud deployment of the system with backup. We used the Turnkey Hub implementation and implemented TKLBAM portable backup. Adaptation of the back-end solution(s) was made available to manage specific Linux deployment. The initial instance was based on open source environment allowing to be openly transferred to other cloud providers. WODA is still being used for data entry and automatic index file preparation for Google Scholar.

Selected cloud deployment - Amazon Elastic Compute Cloud (EC2) - replaced the University server.

The back-up is also provided as a part of the solution. Backups can be used for creation of new instances at different physical location or just for a restore.

Amazon virtual machines are called 'instances', which may have different hardware configuration that match specific requirements of the server. For an initial set-up one can use Amazon Machine Image (AMI), which is a template that contains the software configuration (e.g. an operating system, an application server, and applications). Amazon cloud services allow for easy monitoring of the performance, providing the benefit of elasticity.

Figure 5
A cloud service interface that provides an overview of instance (top) and a completely web based management of a virtual machine for CumInCAD 2.0



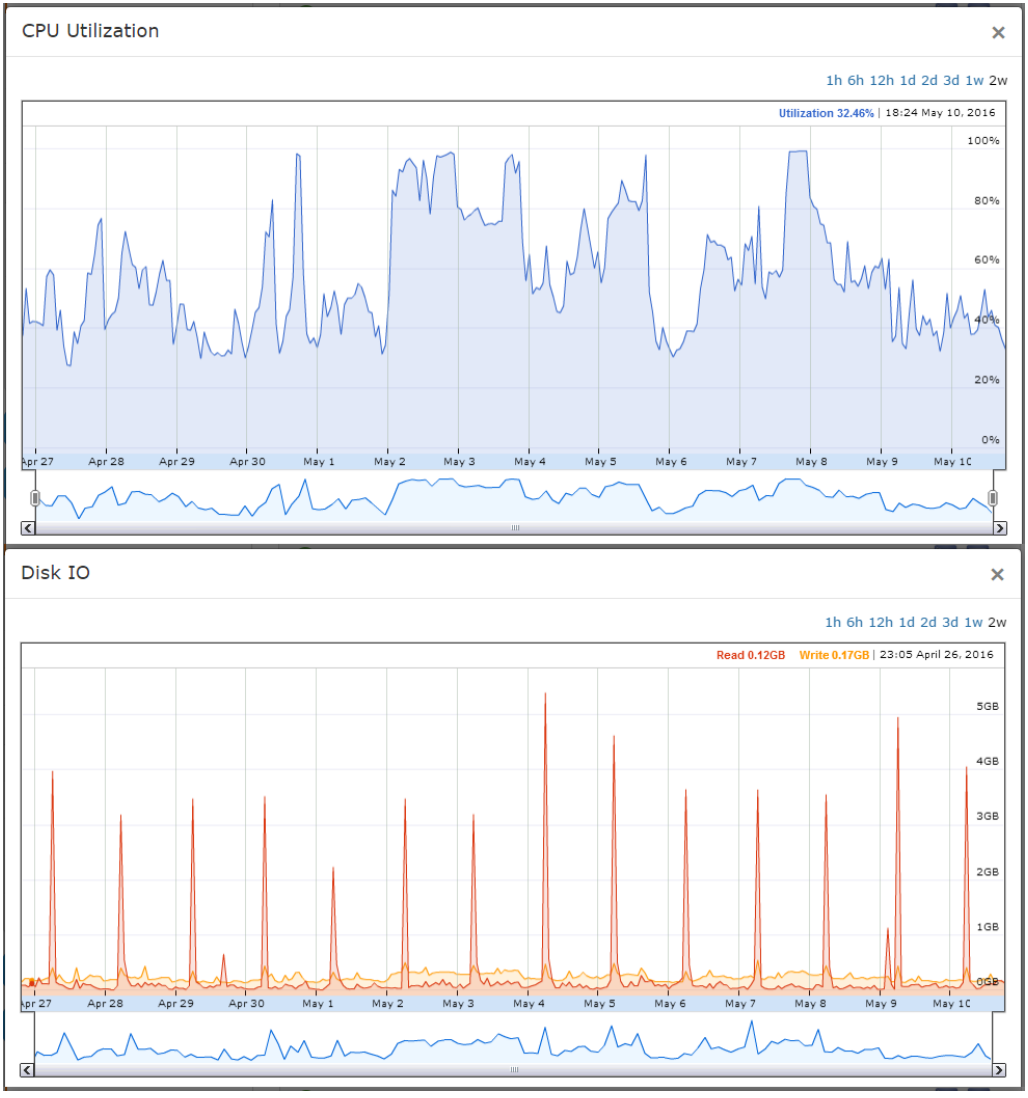
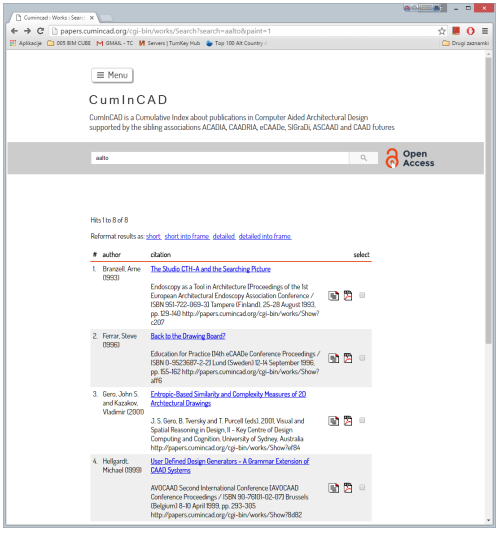


Figure 6
CumInCAD 2.0 selected instance CPU utilization, Disk I/O and network traffic

Testing the ported version on Amazon EC2 was a significant task in order to determine which configuration matched the user requirements of responsiveness and robustness of operations. In this phase we analysed CPU utilization, read/write activities and network traffic (Figure 6).

Figure 7
Redesign of the start page simplified with minimal menu for main features.

Figure 8
Listing of search hits: Example of eCAADe2015-proceedings.



Front-end: Interface redesign

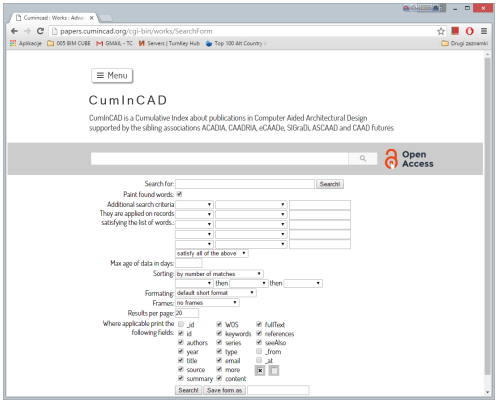
Responsive design considerations guided the re-design of visual representation and the relaunch of the repository website. There were two main principles in the redesign: (1) simplicity and (2) sufficient quality of display on multi-modal devices.

Note that we also moved to a new domain, cumincad.org, in the process of migration, while all existing references to cumincad.scix.net are re-directed to a new domain (papers.cumincad.org).

The focus of delivery by way of the interface is directed towards multiple devices (also for mobile devices - see Figure 9) and above all the interface follows typical search patterns, such as: Certain proceedings of a certain CAAD-association; All published papers of a CAAD-association; Papers of a certain author (personal bibliography), etc.

CONCLUSIONS AND DISCUSSION

The CumInCAD 2.0 cloud deployment proved to be efficient, scalable and independent. The solution is not tied to an institutional IT infrastructure, but is portable to either private or public clouds. We also moved to a new web domain and assured smooth transition with URL rewrite. As CumInCAD is "owned" by the CAAD-associations without being a profit-oriented enterprise intended to generate business, CAAD-associations now have the power to fully control the technical part of the CumInCAD solutions, domain and content.



Finally, help-desk support services to the CumInCAD end-user community and bibliographic data providers (e.g., editors) have been implemented.

Maximising attention to the publication output of individual authors we strengthen the potential impact by making the bibliographic content available for crawling and indexing by Google and other search engines. In this context, we have strengthened the open publishing coverage and implemented advanced indexing and searching. We used the best advantages offered by both worlds, independent developments, portable cloud solutions with integration to the most open commercial developments. For example, a Google team does the daily developments on indexing, searching, ranking, integration, and bibliometrics, while we still do some developments, e.g. Semantic Web, advanced clustering, data mining, along with constant maintenance. We believe that the work invested in this project will result in higher impact through open access supported by search engines and services like Google scholar.

Throughout the implementation we observed the performance of selected configurations of instance and the capacity of selected resources. One of the great advantages of Cloud Deployment is scalability and extensibility.

Cloud interoperability also prevents cloud services and software lock-in. With the help of server usage tracking and support the end-user experience is enhanced and technical solutions can be scaled if needed. In order to assure safe operations in terms of data integrity and backup, we perform daily backups which enable us to restore the systems to a specific date. This way we continue to assure two main goals: safety of valuable bibliographic data and open access.

Figure 9
Listing of search hits: conference series.

Figure 10
Advanced search interface

ACKNOWLEDGEMENT

We gratefully acknowledge the initial work by prof. Žiga Turk and the support of Robert Klinc before the transition, as well as the work conducted by Darko Malić during and after transition.

REFERENCES

[1] http://papers.cumincad.org