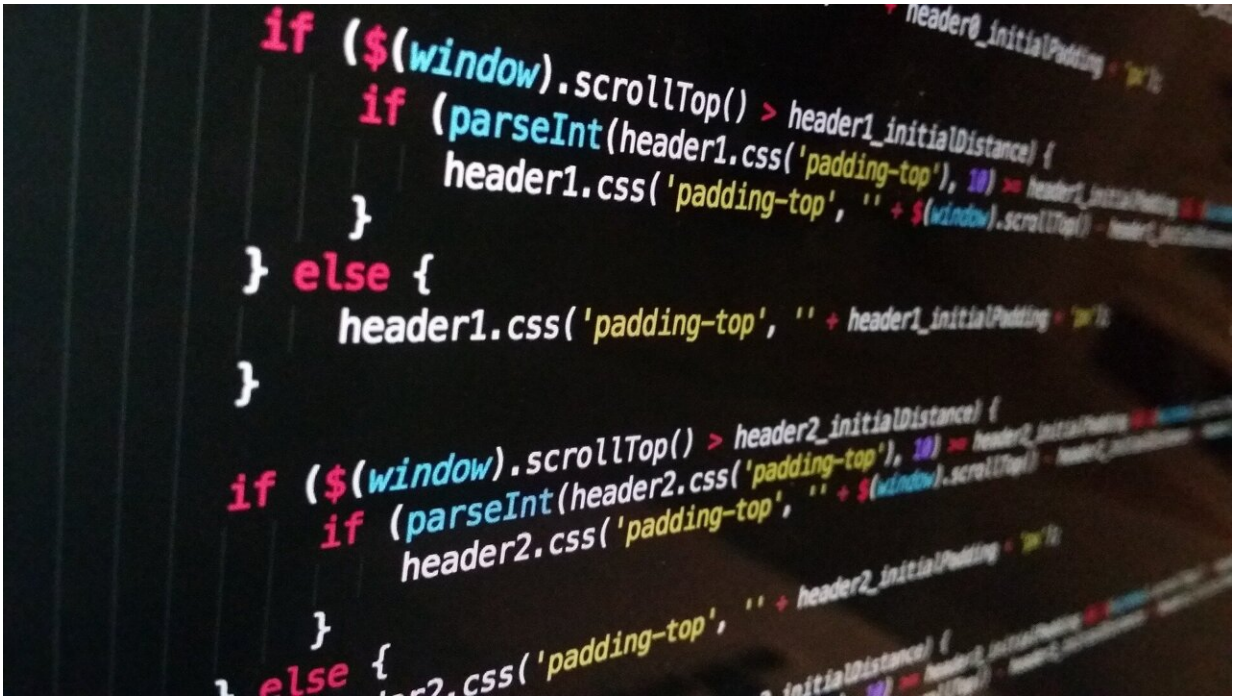
A model for the automatic extraction of content from webs and apps

June 17 2022



Credit: Pixabay/CC0 Public Domain

Content management systems or CMSs are the most popular tool for creating content on the internet. In recent years, they have evolved to become the backbone of an increasingly complex ecosystem of websites, mobile apps and platforms. In order to simplify processes, a team of researchers from the Internet Interdisciplinary Institute (IN3) at the Universitat Oberta de Catalunya (UOC) has developed an open-source

model to automate the extraction of content from CMSs. Their associated research is published in *Research Challenges in Information Science*.

The open-source model is a fully functional scientific prototype that makes it possible to extract the data structure and libraries of each CMS and create a piece of software that acts as an intermediary between the content and the so-called front-end (the final application used by the user). This entire process is done automatically, making it an error-free and scalable solution, since it can be repeated multiple times without increasing its cost.

The importance of CMSs in the online world

Content management systems (CMSs) are behind more than 60% of pages currently available online. Systems such as WordPress, Joomla and Drupal have become popular mainly because they provide a simple user experience, which has allowed all kinds of non-technical users to become part of the online content creation chain.

"Over the last four or five years, these systems have been providing information not only to browsers, but also to mobile apps. CMSs have application programming interfaces (APIs), with which mobile apps communicate to extract content," explained Joan Giner Miguélez, a student on the doctoral program in Network and Information Technologies with the Systems, Software and Models Research Lab (SOM Research Lab) group and lead author of the study that outlines the new model. "These systems, which are known as headless CMSs, allow content, created in a simple way, to be consumed later on different platforms."

CMSs have therefore become a large container of content and data used by each application or platform. This has simplified a lot of processes

but has also added complexities in terms of development that are particularly evident for organizations that manage a high volume of content and platforms. It is increasingly common for the creation of a new mobile app to involve complex development work, and these tasks are simplified by the model designed by the IN3 researchers.

"Imagine a large content company that manages over a thousand websites and apps and wants to make a new mobile app that displays products from each of those websites. If they want to develop the connectors between each website and the application, the work would be immense and resource intensive. It is not scalable," added Joan Giner. "If the APIs are already in a standard format, why can't we also make a content extractor that reads and understands the APIs, represents them in a standard way, and generates the connector to send the information to the new mobile app automatically?"

Automating the extraction of content from CMSs

The model developed by Giner—together with his research partners Abel Gómez and Jordi Cabot, ICREA researcher and leader of the SOM Research Lab—greatly simplifies the development process of a new application and, in turn, results in significant savings in terms of time and resources. The process, which has been developed thanks to funding from the European projects AIDOaRT and TRANSACT, aims to extract and represent the CMS model in a clear and automatic way to make it easier to use as a source of information. In addition, the IN3 researchers' technological proposal aims to generate the code that will act as a link between the CMS and the development of new applications.

To achieve this, the first step is to give the tool the address and login information for the CMS. Once logged in, it reads the API, understands it and uses a reverse engineering process to represent the structure and content libraries of the CMS in a standard way. Based on this, it

automatically generates the connector code through which the CMS and the new mobile app being developed will communicate.

"It is a way of standardizing the process between the CMS and the final application," highlighted Joan Giner. "Its biggest advantage is, in fact, standardization itself. We're talking about a process that is frequently repeated in organizations that manage content; a process that, each time it is performed, involves setting up a specific development team that requires expenditure on a series of resources and that, in addition, can generate errors. Through automation, everything is simplified and becomes more scalable."

As such, this model for automating CMS extractions focuses on scalability, since once the outline and code of the CMS has been created, this can be reused as many times as necessary and integrated into future development projects at no additional cost.

The researchers also point out that it is an automatic model that creates libraries of error-free content, whereas, if the work is done manually, developers can always make a mistake in a line of code.

"Content management systems are a major source of content on the internet. We are making it possible to standardize access to CMSs, just as access to databases was standardized in the past," concluded Joan Giner. "Moving forward, this model could even be used to turn CMSs into a new source of data for training artificial intelligence systems."

More information: Joan Giner-Miguelez et al, Enabling Content Management Systems as an Information Source in Model-Driven Projects, *Research Challenges in Information Science* (2022). DOI: 10.1007/978-3-031-05760-1_30