Motivation and goals

The University of Porto (UP) is a large organization, with dozens of organic units, each one publishing independently several articles per day. This hinders information consumption and potentiates content duplication.

We aim to minimize both these problems by:
• Collecting, storing and consolidating all the information on a database.
• Developing a centralized access point to all the published articles.

Methodology

The data collection process started on June 2012. A database stores all the articles published through the RSS channels available on the SIGARRA platform of 15 UP organic units. Associated data, like the topic, channel and source the article was published under is also stored.

So far we’ve collected more than 6,000 articles published under 56 distinct topics through 108 channels from 15 sources.

Analysis

Between June 1st 2012 and January 31st 2013, we analyzed the information collected to understand the volume and dynamics of the data published.

Sources

Almost half (49%) of all the 5755 articles published during the period in question belonged to five sources: FEUP, FMUP, FLUP, FCUP and ICBAS. The five least productive sources - FMDUP, FADEUP, FCNAUP, REITORIA and FFUP - accounted for about 13% of all published articles.

Topics

The number of topics identified (56) doesn’t match the number of channels (108) because there are several repeated ones. About 71% of all articles were published under 10 topics while the bottom 10 topics didn’t even achieve 1%.

Articles published by month and by hour

November 2012 was the month with the most articles published (982), while August had only 431. The majority of the sources showed similar behavior.

As for the hourly data, we identified a main publishing period that goes from 06:00 to 21:00, with some residual publishing activity outside this period.

Future work

Based on the data collected and subsequent analysis, we are currently working on the design of the interface for the web-based system. We already have an initial prototype online.