**SERVICE DESCRIPTION**

**CALL No. 1/2015**

**SERVICE PROCUREMENT for**

**WEB APPLICATION FOR ARCHIVING MEDIA CONTENTS**

1.1 **Web application for archiving media contents**

The primary purpose of the web application for archiving media contents is to enable the creation of snapshots of webpages and storing them on a secure server with a precise timestamp. The creation of the snapshots should enable the documentation of the content of a particular webpage (including its media content items) and its appearance at a particular instance in time, in two forms:

1. A text (HTML) and images record, to be saved in a database;
2. A snapshot/screenshot, i.e. the appearance of the webpage in a graphical format (JPG and/or PNG)

The development of this application will provide an additional tool for implementation of the Media Fact-Checking Service Methodology, which requires the journalists that work as peer-reviewers to make a snapshot (image) of the currently reviewed article. The reviews also contain direct links to these articles, providing readers a chance to witness for themselves the issues addressed within, but the snapshots are necessary due to the fact that media may change or remove their contents over time. This frequently happens with questionable articles, containing errors or intentional manipulations. Including such snapshots provides evidence supporting the claims of the reviews, to be used by the readers, and as provisional safeguard in case of defamation suits. In this way, the Media Fact-Checking Service is supplementing the public record about the media manipulations.

The web application for archiving media contents will be stored on a dedicated hosting server outside of Macedonia and will provide space for storing the webpage contents selected by its users in two ways:

* Manually - based on the procedure already implemented during the creation of the reviews, the application users will be able to make snapshots by simply providing the web address (URL) of the particular article. Such snapshots stored by the application will be publicly available and any internet user will be able to refer to them by a link, and also to embed the image within their own content. These features will speed up the process of making the reviews, but will also serve the general public, both as a practical tool, and as a way to draw their attention towards other content of the Service, providing an opportunity to engage in a discussion about the contents of the Service.
* Automatically – press clipping of important media websites. Media Fact-Checking Service’s team will be able to instruct the application to regularly and periodically harvest the new content from a select list of websites and store it on the server.

Several snapshots of each article will be made at predetermined intervals in order to account for changes in the article over time. For instance, the application will check for new contents every few minutes, and if it identifies a new published article it will make a snapshot at that moment, and additional snapshots several hours later, the day after and several days later.

The development of the two types of functions of the web application will enable it to serve as a Macedonian version of archive.org, documenting articles and providing them as a public record from the media production, useful to journalists and the general public. Each snapshot will be connected with a timestamp – information about the date and time of the article’s creation, so any internet user will be able to use this information to show how the contents of a particular article looked like at a particular moment in time.

**1.2. Web application for comparing published articles and detecting plagiarism**
Plagiarism takes on new dimensions online as content can be copied and used easily. Journalists sometimes copy pieces of their articles straight from press releases or they just copy the entire article from sources with no or only minor changes. While it’s understandable to occasionally quote press releases and use various sources, over-reliance on government or corporate-designed messages; and copying without attribution or without citing sources is a major problem. The issue of “copy-paste journalism” which includes almost blind copying of contents between online media, making them vehicles for distribution of manipulations, spin and errors, has been identified numerous times in the analyses, reviews and journalistic lessons produced by the experts engaged in the Media Fact-Checking Service.

Fortunately, although it is much easier to perform plagiarism online, with the right tools it is also much easier to detect the original source, or at least the different articles in the group of copied content. In order to support the “linking culture” or “attribution culture” and contribute to the enhancement of professional and ethical standards in journalism in Macedonia, the Media Fact Checking Service will establish a tool that will enable the comparation of published articles online from select range of major Macedonian web sites and to show their similarities/differences statistically in an easy to understand visual representation. In addition, it will visually present a chronology of appearances of articles starting from the source (where the article was initially published) up to its last version, while highlighting the percentage of changes.

The development of this tool will require application of the principles of [data mining](https://en.wikipedia.org/wiki/Data_mining) and visualization, performing the following functions of text mining:

1. Automatic search and download of articles from web pages of Macedonian online news media, through direct connection with a web-application for archiving contents from the media from point 1.1.

2. Automatic identification of relevant information (text, keywords, dates) from these downloaded documents for further content analysis

3. Creation of an online archive storing data about changes to processed posts in a specific time interval

4. Analysis of trends (e.g., such as trending words, events etc.) visually represented in a simple and understandable manner.

5. Analysis of similarities between articles (e.g., determining if a media is plagiarizing articles from another media, with a percentage of discrepancy and sorting similar articles by date and hour of publication) visually represented in a simple and understandable manner.

6. Other automatic analyses that will be considered as an added value to the system, such as determining bias in media using sentiment analysis