2. H4D2 2013 Project: **Processing large amounts of internet data during an emergency**

**Proposed by** [**ACAPS**](http://www.acaps.org/tp:/)[**.**](http://www.acaps.org/tp:/)

**The Issue**

When doing assessments and data monitoring, there’s a need to handle and process a lot of data located in the internet, sometimes under high time pressure. For example the first phase of humanitarian assessment requires using large amount of data to inform decision makers within 72 hours or less from onset of a disaster. In these situations Data managers collate information from media websites, humanitarian forums, government statistics websites or coordination platform and analyse this information in near real time in order to inform decision makers and allocation of resources.  In many way, this activity is similar to web scrapping activity, for selected paragraph or pieces of text located in various places and coming from various sources.

However, the unpredictability of data type, volume, location and velocity means we cannot fully rely on computers and automatization to compile, sort and make use of data. A human brain will always be necessary to select relevant pieces of information and discard less important ones, unreliable information or redundant sources. Rather than harvesting the web using programming and code, an assisted data capture would advantageously assist the researcher in his quest for relevant information and assist him in managing and storing information.

There are currently several softwares that are close to this solution, or build on same principle in large ,e.g. Citavi and Evernote, scrapbook. However, their scope is either too extensive or too limited and usability too complex for the larger amount of data handled for assessments.  To give idea of what sort of data is used and what sort of end product it supports, attached are examples of one spreadsheet filled by hand (copy paste) for the Syria crisis and the type of humanitarian report that is produced once the information is being analyzed.

The desired software solution should have the following functionalities:

* The basic idea is software which works along the lines of Scraper Software, i.e. possibility through human-software interaction to select text, identify attributes, click and save into the database.
* The data used is fetched from various sources and in various formats (HTML, .doc, .pdf, .xls, etc). One important element is hence to have a software that manages to mainstream captured data from various sources, e.g. from various web browsers as well as from native documents stored locally in computer.
* In order to make more effective the data management, a key element is to be able to quickly categorize and sort data into pre-defined dimensions and measure categories. This would provide quantifiable and comparable data at a first step of data fetching. The software should allow the definition of categories and dictionaries in which to choose the relevant “tags” to attribute to each piece of information collected.
* In addition, the software should preferably allow collaborative aspects, i.e. open-source of sorts where several users can contribute and add data into one database (i.e google spreadsheet). This is a key element when time pressure is high, both for sake of triangulation of data as well as quantity.
* Lastly, the software should be able to export acquired data to different formats (.csv, .xls).

Resources:

1. Regional analysis Syria [here](http://h4d2.eu/wp-content/uploads/2013/03/Regional-Analysis-Syria-Part-I-Syria-February-2013.pdf)

2. 130306 DNU aggregated  [here](http://h4d2.eu/wp-content/uploads/2013/03/130306-DNU-aggregated.xlsx)

The ACAPS team, [www.acaps.org](http://www.acaps.org)