“Feasibility study on using automated technologies to support policy-making”

(SMART 2013 – 0024)

Abstract and Executive Summary

Barcelona, 11 June 2014

A study prepared for the European Commission

This study was carried out for the European Commission by

Authors:

Open Evidence:
David Osimo
Fabrizio Smith
Marcello Verona
Katarzyna Szkuta

IES VUB:
Jamal Shahin
Trisha Meyer

DISCLAIMER

By the European Commission, Directorate-General of Communications Networks, Content & Technology.

The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the Commission. The Commission does not guarantee the accuracy of the data included in this study. Neither the Commission nor any person acting on the Commission's behalf may be held responsible for the use which may be made of the information contained therein.

Reproduction is authorised provided the source is acknowledged.

ISBN number
DOI: number

Copyright © 2014 – European Union. All rights reserved. Certain parts are licensed under conditions to the EU.

Abstract

The feasibility study provides a full review of the technological solutions needed for the deployment of the Global Internet Policy Observatory platform. Furthermore it proposes recommendations for its internal governance / management framework. The final analysis resulted in the following set of recommended technological solutions: Nutch and Alchemy for automatic collection of online content, Calais and Alchemy for automatic analysis and classification of content, Mongo DB and Solr for storage of the processed contents, Tableau and D3.js for data visualization through dashboards, and Disqus and Drupal DAA module for user feedback. Concerning the legal framework, the study recommends that GIPO be attached to an existing institution or body. Concerning the management and decision-making structures, it proposes a dual system of advisory bodies - the 'top down'/'formalised' Global Liaison Board and Technology Guidance Board as well as the 'bottom up' Regional Representation Committees and the Temporary Technical Working Groups. It is also recommended that the ex post content moderation would be monitored by the Regional Representation Committees and implemented by the staff attached to the platform. The study team identifies a prioritisation of the steps necessary to reach the end stage of the proposed internal governance of the platform.

Executive summary

The **Global Internet Policy Observatory** is expected to **address the main challenges of the multi-stakeholder governance of the Internet**: a combination of topic complexity, information overload and fragmentation of information between policy silos and in different institutional levels.

The **goal o**f the GIPO initiative is to **make information about Internet-related policies** (including Internet Governance matters) **accessible to stakeholders** are **not currently taking part in the debate**, in addition to **those** who are **already engaged**.

The **feasibility study** aimed at **providing a full review of the technological solutions** needed for the deployment of such a platform, as well as **proposing recommendations** for its **internal governance / management framework**.

It should be, however, underlined that the study **does not aim** at **developing the design, functionalities and architecture** of the GIPO online platform.

The **technologies recommended** by the study are meant to enable **the following functionalities**:

- F1: **automatic collection of online content**
- F2: **automatic analysis and classification of content** (including events and identification of commonalities)
- F3: **storage of the processed content** (together with the output of the analysis), in order to allow the user to search and retrieve relevant information
- F4: **visualization** of the data through "**dashboards**"
- F5: **evaluation of the information** in such a way as to **allow users** of the platform to express their **assessment of the quality** of such information.

The technological assessment was carried out **in four steps**, a **mapping phase** that provided 30 technological solutions, an **initial "relevance" assessment phase** that enabled the identification of five solutions for each of the required functionalities, an **in-depth assessment** which recommended two solutions for each of the functionalities, and a **final analysis** of those two solutions based on the implementation of an online demo.

The **final analysis resulted in the following set of recommended technological solutions.** The results are summarised in the table below.

Functionality 1 – automatic collection of online content- Nutch & Alchemy
Both tools provide a **solid and scalable environment to crawl web contents and documental resources**, extracting the main informative content. Nutch offers a **more scalable solution** for **web resource fetching** and **directly supports documental resources encoded in the main formats** (e.g., docx, pdf). It also includes a **data storage layer,** which can be seamlessly combined with a number of NoSQL storage solutions. It only requires a **list of initial "seed" URLs** to start the crawling process. Only **limited configuration efforts** are required, also in case of deployment in a computer cluster since Nutch smoothly integrates with the Apache Hadoop Framework. Alchemy requires **higher set-up efforts**. Furthermore, **the keyword extraction capabilities** of Alchemy can be used to **improve the quality** of the collected data, by selecting contents according to domain-specific filters (i.e. sets of relevant keywords). However, performing such operations requires a great amount of invocations to the Alchemy web server, making the overall system **susceptible to network speed and server load, and limiting the scalability**.
Functionality 2 - automatic analysis and classification of content - Calais and Alchemy
Both tools are widely adopted and highly effective solutions for **semantic analysis and the enrichment of unstructured information**. Their functionalities are in both cases offered as cloud-hosted services, reducing **the efforts required for the set-up, maintenance and integration** into **external applications** to a minimal level. Alchemy offers, through a single access point, a **wider set of semantic analysis** functionalities than Calais. In addition, it offers a **more advanced multi-language support, crawling capabilities and sentiment analysis**. On the other hand, Calais is focused on **the extraction of statements regarding *events*, *facts* and *entities*** (e.g., persons, organizations, locations). In both cases, the **produced semantic annotation** can be **effectively exploited to identify commonalities and relations among different pieces of content**. **Additional efforts** should be dedicated to the **construction of background terminological models**, specific to the domain at hand, **in order to implement a context-aware interpretation of the produced semantic meta-data** (which are intended to be domain-independent), thus improving the **overall analysis accuracy**.
Functinality - storage of the processed content - 3 Mongo DB and Solr
Both tools provide a solid and scalable document-oriented storage and query system. MongoDB is a **very flexible NoSQL document database** characterized by **high vertical/horizontal scalability**. It offers a **powerful query suppor**t to retrieve documents from a collection on the basis of multiple criteria. Solr is a reference framework for keyword-based indexing and content search. In addition to NoSQL data storage and querying facilities, Solr provides relevant **features** including **full-text search, faceted search, rich document handling and geospatial search**. For the fine-tuning, both solutions offer **many optimization techniques**, and a **direct support for the deployment** across multiple machines, addressing cloud or computer cluster installations.
Functionality 4 – visualisation of data through dashboards - Tableau and D3.js

Both solutions are **widely adopted and highly effective tools for (big) data visualization and analytics**. Both solutions allow for the combination of different visualization types in **customized dashboards**, giving the user the possibility to select different "views" over the same datasets with a very limited effort and without the need of particular technical skills or training.

Tableau is a **commercial product** that **requires less initial efforts**, for installation and connection to the data sources (notably it offers data integration features. Further the deployment of customized visualization takes place through a graphical platform (graphical user interface) intended to be **accessible to non-technical users** (at least, for what concerns the basic features).

On the other hand, the **open-source library D3.js** requires **greater technical skills and implementation efforts**, which pays back in terms of **absence of licensing costs**, **flexibility, possibility of integration with third parties visualizations**, and **standard compliance** (fully W3C-compliance, making use of the widely implemented SVG, JavaScript, HTML5, and CSS3 standards).

Functionality 5 – users' evaluation of content Disqus and Drupal DAA module

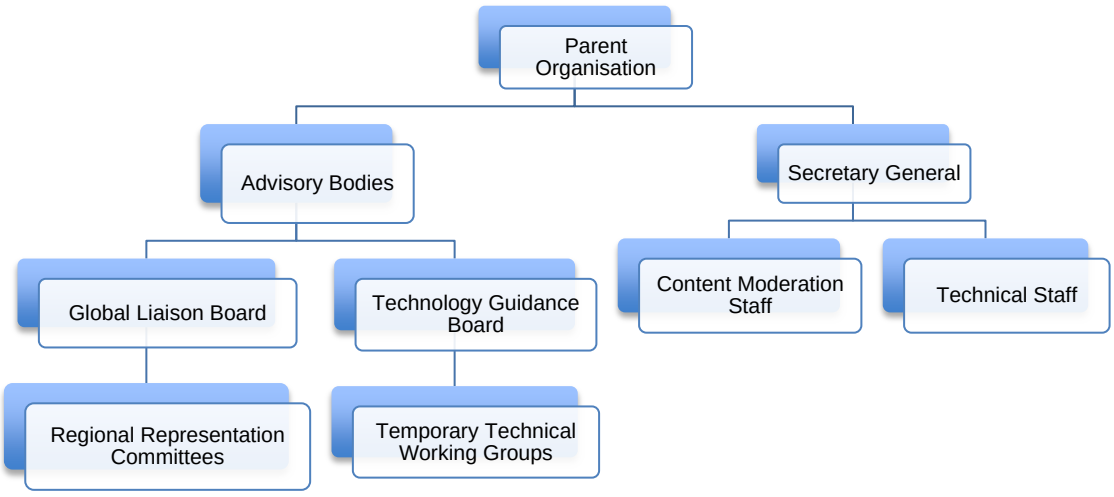
The choice here is not really between two technologies, but between two models. Fr**ee software as a service**, such as Disqus, enables **very easy deployment and seamless native integration with social media**. However, it **implies limited control over data** and **little possibility to customize**. Moreover users might be reluctant to give their data to a commercial third party platform.

Drupal DAA module, on the other hand, **requires more initial effort** in writing and also in maintaining it, but **provides full flexibility and ownership of data**, which could be reassuring for users.

The study team also provided a set of recommendations for an **internal governance / management framework.** The recommendations provided are **based on a review of literature**, an **analysis of previous experiences and a limited set of interviews**. These set out to make recommendations and a specific proposal for an internal structure for organising the platform's non-technical resources.

The process of determining the GIPO institutional framework will be a dynamic one, and – given the contentiousness of the political discussions around global internet governance – should not be restrained by predesigned recommendations and models but rather inspired by the study team's proposal. The study team **strongly advocates** close **monitoring of the evolution** of the platform's governance framework in order to ensure that key values are maintained.

According to the study team, the GIPO platform should take the following structure into consideration during its evolution.



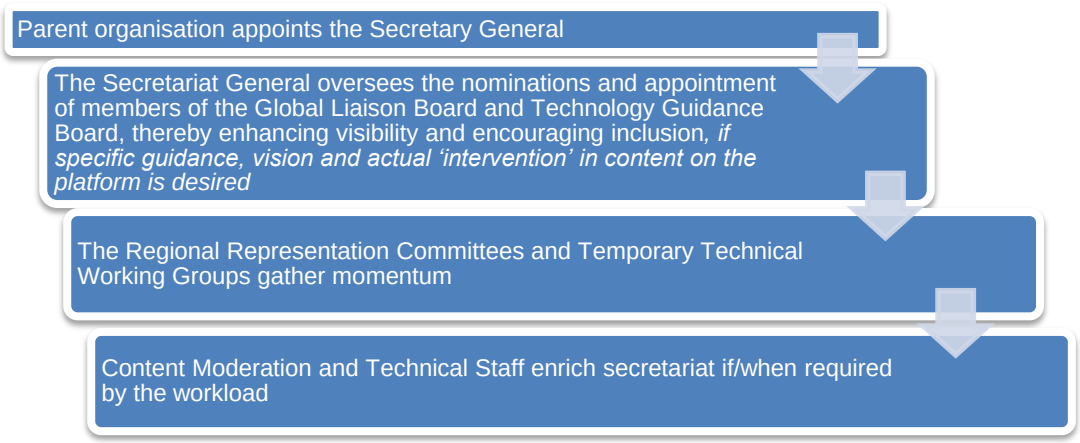
Concerning the **legal framework**, the study team recommends that GIPO be **attached** to an **existing institution or body**, such as the IGF Secretariat, based in Geneva. Other instances that could be engaged as the parent (hosting) organisation include the W3C, or the OECD.

Concerning **management and decision-making structures**, the study team proposes the following. In order to ensure both stability and community engagement in the platform, **a dual system of advisory bodies** should be created. The **'top down'/'formalised' Global Liaison Board and Technology Guidance Board** should be composed of **a limited number of established experts** in their fields that are appointed by the platform's 'parent organization.' (hosting organisation). Their appointment would be non-remunerated and for a fixed term. The '**bottom up' Regional Representation Committees** and the **Temporary Technical Working Groups** should be **flexible bodies composed of self-appointed volunteers** who are assembled in a much looser framework.

The **Global Liaison Board** should be established if **specific guidance and vision** in terms of direction, community building through reputation of specific key individuals, and actual 'intervention' in content on the platform **is desired**. In the study team view, this organisational body is necessary to promote **inclusiveness and relevance of the platform**.

The **key tasks of the main actors** engaged in the GIPO platform are **highlighted in the table** below. **Ex post content moderation** would be monitored by the **Regional Representation Committees** (with community members or users also able to make recommendations according to a flexible rating system) and **implemented by the staff** attached to the platform. The **Regional Representation Committees** would provide **content sources**; the **Global Liaison Board** could provide **specific content, such as content they consider to be especially important, or curation on specific issues, which may be grouped into focus areas**.

The internal governance processes and structures of the platform will naturally evolve in time. Thus, the study team **identifies a prioritisation of the steps** necessary to **reach the end stage** of the proposed internal governance of the platform.



Given that the platform will need to satisfy a diverse range of stakeholder needs and desires, the resulting institutional framework may not – in the end – resemble the proposed solution: as with most developments in institutions, there is an organic evolutionary process to be undertaken. However, we feel that the vision of the institutional framework proposed herein is a useful starting point.

Based on the technological assessment, the study team concludes that **existing technological options are available, mature and would incur limited costs**, since they are mostly based on **open source software**. The integration of the **technologies requires development work.** Yet, during the feasibility study phase, it has been possible to have a working demo integrating the different technologies, within the resource and time constraints of this study.

At the same time, while technologies are mature enough, **the implementation is challenging**. We have not been able to identify a similar example of technological solutions applied to policy observatories, as policy remains a complex theme with highly contextual information and is generally addressed with extensive human effort. In this sense, **GIPO is a pioneer project.**

DOI: 10.2759/94701 ISBN 978-92-79-28122-8