SOCIETIC
SOCiety as Infrastructure for E-Science via technology, innovation and creativity

Deliverable no.	D4.1.
Deliverable name	Semantics map application report
Dissemination level	PU
WP no.	4
WP name	Use of the infrastructure. Applications and Creativity
Date	17/04/2013
Date of delivery	31/3/2013
Actual date of delivery	17/4/2013
Status	Final
Author(s)	Fermin Serrano, Francisco Brasileiro, Jose Antonio Leal de Farias
Reviewer (s)	Nazareno Andrade, Javier Borge, Eduardo Lostal

SOCIETIC is supported by the European Commission under Contract Number: RI-312902

European Commission

Change log

Version	Date	Author/Editor	Reason for change / issue
1	12/4/2013	Fermin Serrano	Creation
2	15/4/2013	Fermin Serrano, Francisco Brasileiro, Jose Antonio Leal de Farias	Review
3	17/4/2013	Fermin Serrano	Final

Disclaimer

This document is the property of the Societic Consortium. This document may not be copied, reproduced, or modified in the whole or in the part for any purpose without written permission from the Societic coordinator with acceptance of the Project Consortium. This publication was completed with the support of the European Commission under the 7th Framework Programme. The contents of this publication do not necessarily reflect the Commission's own position.

Table of Contents

1. Summary

This deliverable compiles scientific, technological and social aspects regarding the semantic-maps application, covering analysis, deployment and initial results.

The application is in production, despite its visual aspect is not finished yet, that is the reason to mark this application as beta. Firstly, we developed a 'mockup' version of the Semantics App for evaluation, which was used to evaluate the user interface and the basic functionality needed to experiment. During this period, experiment requiremetns have been improved and modified, and different versions of the database were sent for testing. Eventually, a production candidate version was created for Pybossa, which is now in the testing stage.

All versions of the application have been developed using HTML5 + JavaScript on the client side, and Python on the server side.

Initially the WP4 plan was designed thinking in having deployed Semantic maps application first (deliverable month 6) and Cells application later (deliverable month 10). We decided to speed up the development and publication of the Cells application taking advantage of the academic year so that secondary schools students from Spain can join the experiments. Initial evaluations, results and feedback have been obtained already. In the 2013-2014 academic year, we plan to launch the experiment at European level. Thus, further information will be included in the deliverable focused on that experiment.

2. Introduction

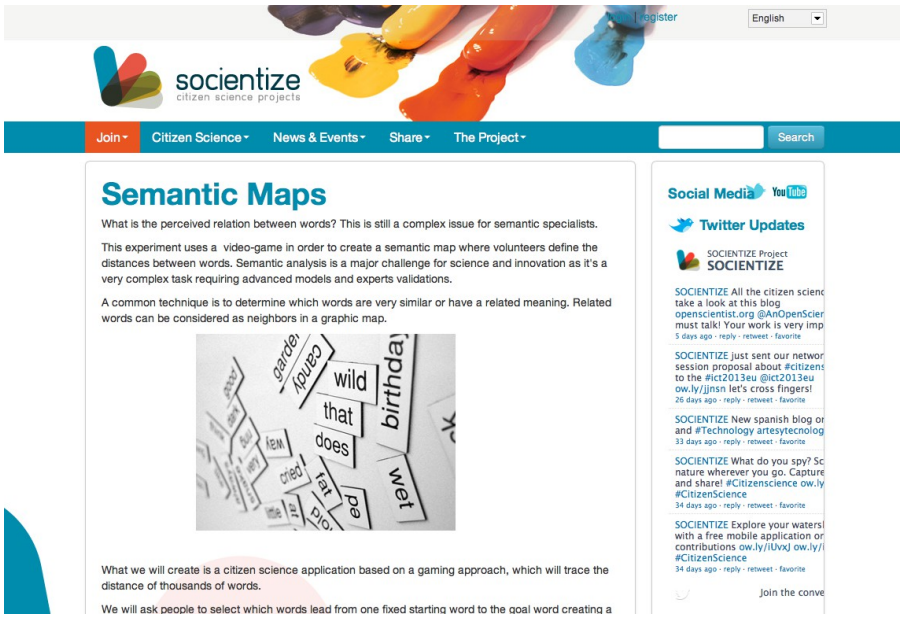
During the last ten years several studies have appeared regarding language complexity. Research on this issue began soon after the burst of a new movement of interest and research in the study of complex networks, i.e., networks whose structure is irregular, complex and dynamically evolving in time. Beyond the study of language structure (a static view of the main descriptors: clustering, average degree, etc.) which motivated the initial efforts, research has slowly shifted towards an understanding of language dynamics (network navigation or emergence of semantic meaning, for instance). This is relevant not only to understand the cognitive mechanisms underlying language usage, but also interesting in relation to search engines (which seek increasingly precise and efficient retrieval strategies) or neuromarketing (which aim to learn cognitive responses from consumers, when faced to certain information), among others. It is clear then, that the number of interesting questions in the field is vast; the main problem being that of empirical validation and data collection. In this context, new avenues of research are opening up as we gain access to a large fraction of the population through ICTs, who provide scholars with an unprecedented amount of data.

Focusing on semantic navigation, we know from past work that humans are very good at navigation even in foreign environments such as the World Wide Web. The question is now whether individuals can efficiently navigate semantic space, and what heuristics --if any-- they use. In practice, we give individuals a start location in the semantic network and ask them to build chains

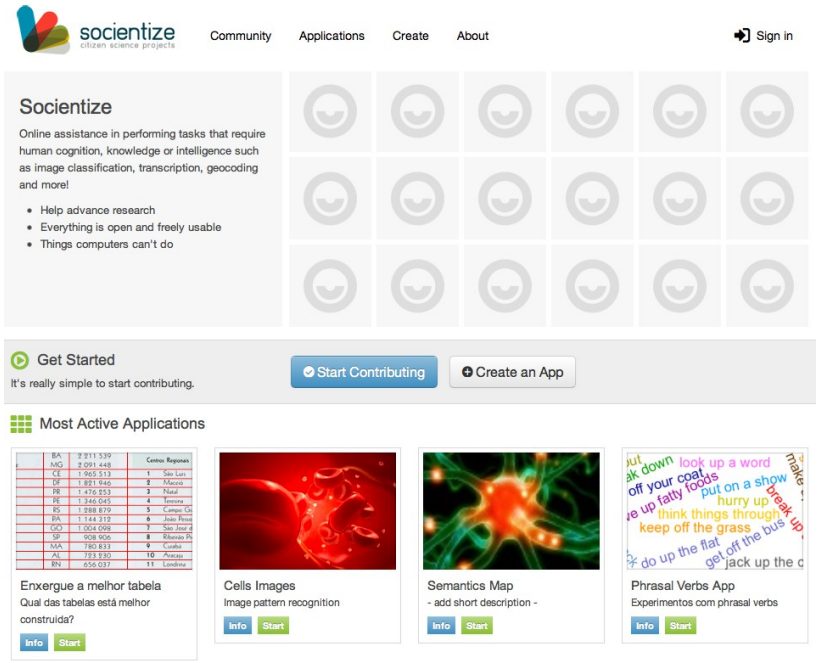
of associates to get to a target location. This data provides us with the decisions of individuals which can in turn tell us about the underlying navigation process and semantic environment. Specifically, we can compare empirical chains (those obtained from participants) to synthetic ones, such as the shortest path which can be regarded as the baseline optimal navigation, similarity paths which exploits computationally estimated word-word similarities or random paths (our navigational "null model"). The outcome of this systematic comparison should pave the way to a reverse-engineering effort that uncovers which factors constraint human semantic navigation.

The experiment, under the form of a game, can be presented in many ways. As a starting point, we present a minimally biased version of the game. This means that we try not to condition the participant at all (the interface uses words with the same size, color, etc. And items are presented in a randomized fashion, so as to avoid "position bias", etc.). As data collection grows, we intend to investigate how implicit information affects heuristics and performance, i.e. whether participants change the way they navigate when these (or other) clues are available. This can be attained under a variety of ways, such as changing the word's size as a function of its centrality in the network.

The experiment is accesible from the www.socientize.eu website and also from the PyBossa specific website.



Picture 1. Screenshot of the web page of the experiment in the Societic website (www.socientize.eu)



Picture 2. Screenshot of the PyBossa web page of the experiment in the Societic website (pybossa.socientize.eu)

3. Application specifications

3.1. Storage

There is a MySQL database containing the list of words and the links. The database name is semantics. It contains two tables:

- words: it contains an id for each word and the word itself. (id int, word varchar(20))
- links: it contains the directed link and the weight (normalized) of it. (id1, id2, weight)

3.2. Description of the application

Three words are presented to the final user. The first one is the source and the third one is the target. Second one, in the middle, is the current word. A cloud of words around the current one is presented also to the final user. These words are the ones that are linked to the source word in the links table (remember that this table presents directional links, that is, a link between a word A to word B does not mean the link from word B to word A exists).

User can select one word of the cloud. Then the selected word is placed in the center of the cloud, the old cloud disappears and a new cloud of words (the linked ones to this new word in the links table) is presented.

When user clicks on the target word, game ends and a new pair of words is shown to the user,

starting again.

As above mentioned, two versions of the application are prepared. Both versions share several common features since the core of the application is the same. Differences are found in the presentation of both of them, regarding the amount of implicit information to be displayed.

Common features:

- Source and target words will be presented to the user at the middle part of the screen: source word at the middle-left, target word at the middle-right. As the game advances, path followed by the user is built. That path is displayed below the source and target ones: each selected word by the user is added to the path, so it is possible to keep track of what words have been chosen.

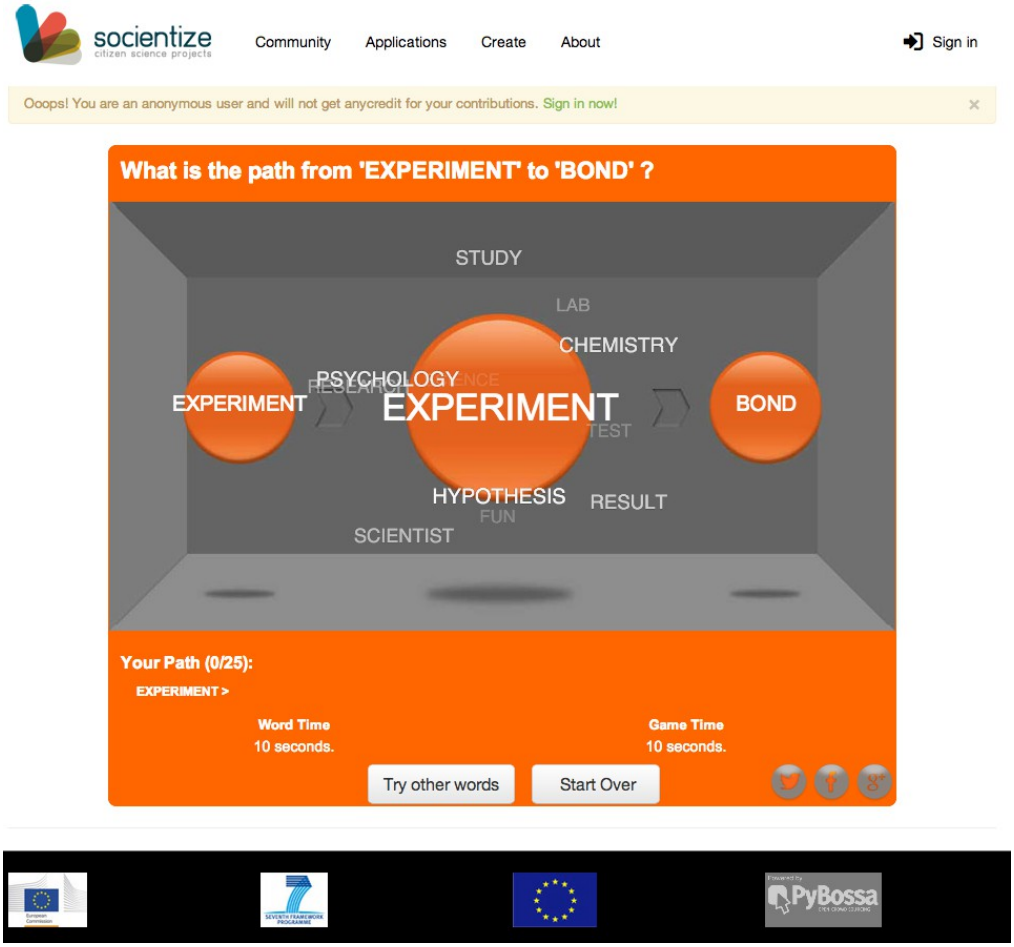
- The word that the user is currently evaluating should appear in the center of the screen, and the set of words among which he/she has to choose from will appear surrounding it (forming a circle around the current word).

Differences:

- In a first stage, the game will provide a “uniform” presentation. That is, size of words will be the same for all possible choices. By doing so, we are preventing that the user takes advantage of some implicit information (like, who is best connected).

- In a second stage, words will have different size according to the grade of the word (the grade of the word is the addition of the incoming and outcoming connections), benefiting the best connected words. More connection, bigger the word. The idea behind this is that the user can learn some heuristic (such as: choosing the best-connected words typically leads to quicker success).

Beside both versions are already implented at the moment of the creation of this document, “uniform” version is the only one being deployed initially, waiting for initial results.



Picture 3. Overview of the application user interface (under improvement)

4. Technical Report

We can divide the project into two major blocks, the "view" and "server".

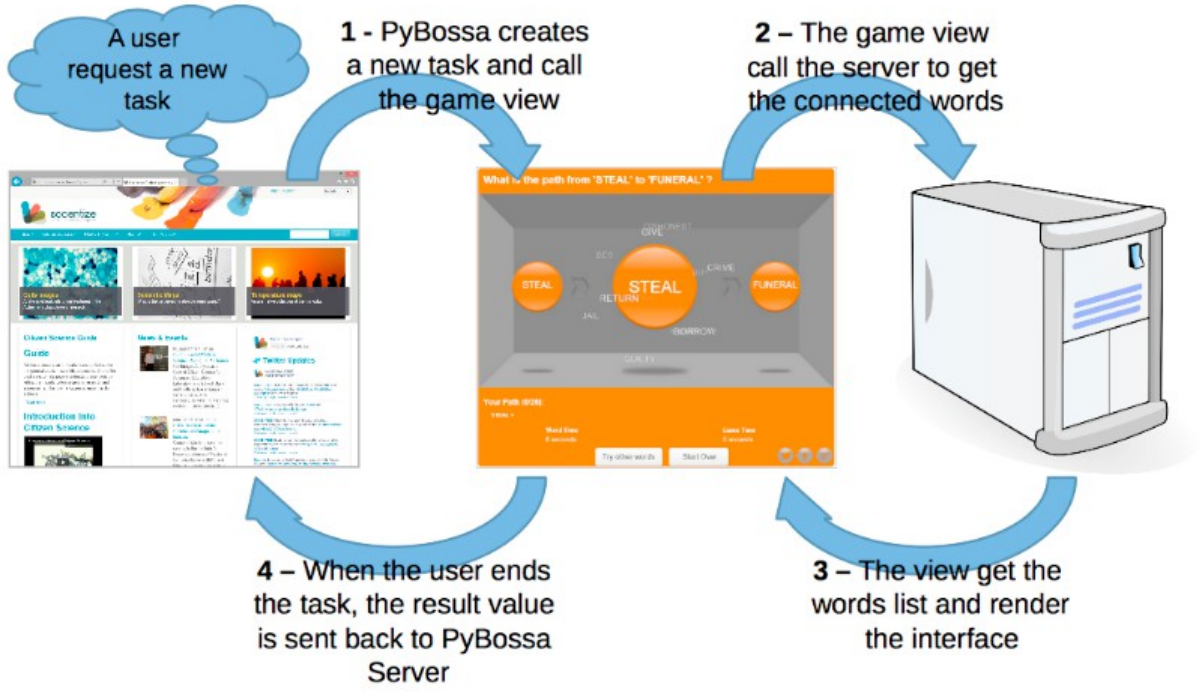
The view is a graphical interface for game shown in the web browser and run entirely in the browser. On the other hand, the server is the application that controls the state of the game and the execution of tasks, and runs at the Pybossa's server machine.

The view was developed in HTML5 and make asynchronous calls (using JavaScript) to the server, sending and receiving only the necessary data to the current state of the game in order to minimize network traffic and provide a consistent user experience, without delays, during the user session.

The server uses a MySQL database server to store all words (nodes) and their vertices (links). It was developed in Python, using the Flask library to respond the asynchronous HTTP requests from the view, sent when it is necessary to obtain the set of words related to the current word, which is then sent back to the words cloud. The server is also responsible for creating the entire application and

its tasks on the PyBossa server.

Next picture shows the data flow between the view and the application server:



Picture 4. Worflow designed for the experiment

All development efforts follow the SCRUMM methodology (with weekly sprints), using the Visual Studio 2012 as the main development tool.

4.1. Features implemented

- Maximum steps to reach the goal word: 25 steps.
- Maximum time between clicks/words (maximum time between a word and the next one can be chosen, that is, maximum time since the last click): 1 minute
- If the game is abandoned or the time limit is exceeded, that game is considered as abandoned and consequently, information about that game is not stored and the replica is not considered as valid.
- If a user doesn't reach the target word because the step limit (25 steps) has been exceeded, the game is considered as a valid replica but not successful. Information in such a case must be stored and user encouraged to keep playing.
- It is not possible to come back, that is, it is not possible to undo a decision
- Besides the target word appears in the current set of words, the game won't finish until the word is clicked. This is, the target word must be clicked in order to finish the game.
- Set a high number of replicas for each task, the idea is to get as many results as possible (from 5 to 500).

- Things to be stored for each game: success, path and times between word selection.
- Add a countdown of the possible steps left out which must be visible for the user.
- Make the words uniform: either use lower case letters or starting with capital letter. the same for the colour since more appealing colours might be benefited if colours change for the words.
- Set of words related to every word will be displayed randomly for every replica/task
- "Reset" button.
- Words already passed through will be able to appear again if they are among the neighbours of the current word.
- Save the number of resets done.
- Share tool via social networks (google+, facebook, twitter)
- Include link self-avoiding algorithm

5. Initial results

As we said, there are many phases regarding different scientific approaches to the problem. Initital phase (alpha-beta testing) is 45% completed. Next phases will require bigger citizen contributions. Outputs and results format are JSON, text-based open standard. One example of output:

[
{"info": "ANYTHING~SAY~WRITE~PEN", "user_id": 4, "task_id": 5151, "created": "2013-03-20T21:51:59.143005", "finish_time": "2013-03-20T21:51:59.143037", "calibration": null, "app_id": 418, "user_ip": null, "timeout": null, "id": 910},
{"info": "ANYTHING~SOMETHING~OBJECT~PENCIL~PEN", "user_id": 2, "task_id": 5151, "created": "2013-03-21T02:34:23.407033", "finish_time": "2013-03-21T02:34:23.407067", "calibration": null, "app_id": 418, "user_ip": null, "timeout": null, "id": 921},
{"info": "ANYTHING~SOMETHING~THING~BOOK~PAPER~PEN", "user_id": null, "task_id": 5151, "created": "2013-03-21T14:18:43.004810", "finish_time": "2013-03-21T14:18:43.004836", "calibration": null, "app_id": 418, "user_ip": "155.210.136.33", "timeout": null, "id": 967},
{"info": "ANYTHING~SOMETHING~OBJECT~PENCIL~PEN", "user_id": null, "task_id": 5151, "created": "2013-03-21T14:23:36.287536", "finish_time": "2013-03-21T14:23:36.287562", "calibration": null, "app_id": 418, "user_ip": "155.210.136.22", "timeout": null, "id": 968},
{"info": "ANYTHING~SOMETHING~OBJECT~PENCIL~PEN", "user_id": null, "task_id": 5151, "created": "2013-03-27T09:33:46.833726", "finish_time": "2013-03-27T09:33:46.833755", "calibration": null, "app_id": 418, "user_ip": "85.251.81.108", "timeout": null, "id": 2995},
{"info": "ANYTHING~SOMETHING~OBJECT~PENCIL~PEN", "user_id": null, "task_id": 5151, "created": "2013-04-05T14:21:44.384285", "finish_time": "2013-04-05T14:21:44.384308", "calibration": null, "app_id": 418, "user_ip": "155.210.136.68", "timeout": null, "id": 3069},
{"info": "ANYTHING~SAY~WORD~PHRASE~SENTENCE~NOUN~GRAMMAR~BOOK~PAPER~PEN", "user_id": null, "task_id": 5151, "created": "2013-03-21T21:01:16.009186", "finish_time": "2013-03-21T21:01:16.009228", "calibration": null, "app_id": 418, "user_ip": "155.210.136.60", "timeout": null, "id": 1554}
]

The whole set of outputs will be published once publications are made by the research team.

8. Conclusion

The application is in production accesible for the general public in the project website, in the URL http://www.socientize.com/?q=eu/content/semantic-maps and http://pybossa.socientize.eu/pybossa/app/Semantics/newtask

Waiting for a better user-interface design, no big publicity has been done. It is also due to the fact that the application of cells images analysis has been also developed during the first 6 moths of the project.

All the test performed with the users are very satisfactory, with great user experience: funny, challenging and competitive.

Further developments and new phases will get larger awareness, major social impact with larger community of users. Those dissemination plans and scientific results will be presented in the following project progress reports.