Deliverable 5.1: User tasks and requirements; data abstractions and operations requirements

Project Acronym	OPENCARE	
Title	Open Participatory Engagement in Collective Awareness for REdesign of Care services	
Project Number	688670	
Work package	WP5 – Data processing for aggregating collective intelligence processes	
Lead Beneficiary	UBx — University of Bordeaux	
Editor(s)	Guy Melançon	*University of Bordeaux*
Reviewer(s)	Guy Melançon Bruno Pinaud *Alberto Cottica*	*University of Bordeaux* *University of Bordeaux* *Edgeryders*
Dissemination Level	Public	
Contractual Delivery Date	31/12/2016	
Actual Delivery Date	27/12/2016	
Version	1.0	
Status		

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688670.

Summary

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688670.

D5.1 White paper: user tasks and requirements; data abstractions and operations requirements

This document reports on activities conducted in order to specify the design and goals of the:

- semi-automated aid to ethnographic coding
- SSNA dashboard environment

Both of these tools are made available through the web browser. A prototype can be accessed at the URL http://164.132.58.138:9000/. This URL is temporary; later, more stable and complete versions shall migrate under the opencare.cc URL and/or the edgeryders.eu portal. For now, our user base being still limited, the two set of tasks and views are presented through a single application.

In this document, we go over the methodology we follow in designing the SSNA dashboard and designing a semi-automated aid to ethnographers using the Open Ethnographer Drupal plug-in.

It is useful to report on the design process at this stage, to help stabilize our design choices. We however adopt an iterative approach (see next section): the design is turned into prototypes that are submitted to users; earlier prototypes are adjusted according to user feedback and re-submitted, improving our design choices and the overall usability of the dashboard.

Designing visual analytics environments

The process of designing the dashboard followed Munzner's four level model (Munzner 2009) for designing visualization systems, later revisited by (Meyer 2012) (Sedlmair 2012). The model promotes a hierarchical approach ordering issues and design questions to be solved before the system is actually implemented.

- Domain Problem and Data Characterization
  - Operations and Data Type Abstraction
    - Visual Encoding and Interaction Design
      - Algorithm Design

Domain Problem and Data Characterization. This crucial step is where designers and/or implementers must learn about the tasks and data of target users. Each domain usually has its own vocabulary for describing its data and problems. Collaborative workshops were organized (see deliverables D1.1 and D1.2) to engage designers with target users, favor the development of a common understanding and common concepts around which user tasks (and ultimately the dashboard) are built.

The primary output of these collaborative sessions is a series of domain questions and actions to carry out on the data. These questions were collected during the workshop in collaborative documents (on the cloud), or reported afterwards in discussion threads. Some of these questions will be listed here.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688670.

Operations and Data Type Abstraction. The collaborative sessions then gave way to an abstraction stage to map problems and data from the vocabulary of the specific domain into a more abstract and generic description using the vocabulary of computer science, and more specifically that of Visual Analytics.

Describing user tasks as operations means putting them as generic actions performed on data (through visual representations and human-machine interactions) (Wehrend Lewis 1990) – cited by (Munzner 2009) and later by (Renoust 2013). As Wehrend and Lewis put it, *designers need to distinguish problems/tasks for which the user's goal in viewing a graphical representation differs*. For instance, a representation may, in some cases be used to *identify* a value or entity, in others to *compare* between entities.

The other aspect of this stage is to transform the raw data into the data types that visualization techniques can address.

Visual Encoding and Interaction Design. Because visual encoding and interaction are mutually interdependent, we consider them together rather than separately. It indeed often is a pair (graphical representation x interaction) that best supports a task. Incidentally, a same graphical representation may thus serve multiple tasks.

Since the seminal and foundational work from (Mackinlay 1986) and (Card et al. 1999), and (Herman 2000) for what concerns network visualization, the design of visual encodings has received a great deal of attention in the literature.

Algorithm Design. This is the lowest level issue that need to be addressed, where we may have to create algorithms to carry out the visual encodings and interaction designs. Obviously, these issues are not necessarily unique to Visual Analytics, and may rely on ideas and pre-existing solutions part of the computer science literature.

The description of Munzner's model may lead to think it is executed in a linear manner, going from domain questions down into the lower levels. We however followed a iterative process, where domain questions are confronted with the available data, and later on when the first visualizations were designed and prototyped.

Domain questions and high-level goals

We list here domain questions and or goals expressed by users during collaborative sessions. There are also a number of discussion threads where tasks/functionalities were suggested and/or discussed (which have been reported in deliverables D1.1 and D1.2)[*].

[*] See for instance, the Masters of Networks 4 thread reporting on early discussions:

https://edgeryders.eu/en/lote5/masters-of-network-4-networks-of-care

Or the after-workshop documentation gathered into a collaborative document (hackpad):

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688670.

Users have different profiles falling into (at least) one of the following categories:

- Community managers
- Larger (non specialist) audience
  - Community managers, and to a lesser extent a larger audience, forms the user base of the SSNA dashboard.
- Ethnographers
  - Ethnographers form the user base for which the Open Ethnographer semi-automated aid is designed.

Community managers and larger audience. The workshop we conducted and discussion we had led users to express high-level requirements.

(*Adapted from discussion threads and/or workshop collaborative documents*)

Roughly speaking, the dahsboard should:

- support the high-level task aiming at helping care professionals or activists as well as patients, and actually anyone interested in contributing to the opencare ecosystem, to *make sense of collective intelligence* as it takes place through online conversations.

More generally, the dashboard should help:

- navigate and explore traces of conversations helping users to *make sense of the different ways in which people engage in online conversations* (mailing lists, forums, social environments like twitter, or other independent platforms).

Community managers may try to:

- observe ("measure"?) the effect (if any) they have *on the way conversations develop* in (a) the opencare community?[†]

At some point, sentiment analysis was mentioned as a possible analytical aid. It however did not receive much support from community managers and was left aside[‡].

---

https://lote5.hackpad.com/SAT-0930-1045-MASTERS-OF-NETWORKS-NETWORKS-OF-CARE-hackathon-for-network-scientists-doctors-and-patients-to-make-sense-o-vxaFSnxaNTg

https://edgeryders.eu/en/lote5-doc/documentation-masters-of-networks-networks-of-care

[†] On this, a paper co-authored by (Cottica, Melançon and Renoust 2017) may be of interest (recently published as part of volume 693 of Springer Studies in Computational Intelligence). An open access version of the paper is accessible on HAL (affiliated with ArXiV).

[‡] See https://edgeryders.eu/en/opencare-research/sentiment-analysis-play-the-game.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688670.

Ethnographers. Sessions were run with ethnographers, and "ethnography-aware" community managers, to identify how the visualization of data could complement and/or support the work of ethnographers.

Ethnographers code content, that is, they go through discussion threads, select pieces of text and associate to it a *code* (also called a "*tag*"). This allows community managers and users to gain a higher-level view on what is taking place within the conversations as a whole. It can also be used to select threads of interest, etc.

(*Excerpts from workshop sessions held in Stockholm, June 2016 and Milano, November 2016*)

Ethnographers have expressed several requirements to help them navigate through all discussion threads and gain a higher-level view of their own work. It turns out that some of these requirements could well be made available to a larger audience as well.

Task set #1

- See how tags cover the conversations being examined.
- From a set of tags, find the most important conversations associated with them.
- Find "rich" conversations; find the most "insightful" post, the one engaging the most people.
- Picture the tags most often associated with a person (through the posts/comments they author).
- Provide feedback on the "level of expertise" of users involved in a discussion thread.
- Tags associated to a given user could be pictured using a tag cloud.

Task set #2

- Distinguish "rich" posts or comments, those having a larger number of associated tags, and presumably being longer posts.
- The notion of a "popular" tag (associated with more content) also came up as being of interest.
- The number of persons involved in a post is an interesting statistics.

Later on, it appeared that tags naturally form into a hierarchy and that making this hierarchy explicit in the navigation could prove useful.

- Make keyword categories and work on these categories to find conversations and users.

Operations and Data Type Abstraction

The data we deal with actually exists and is recorded in a database underlying the Drupal website deployed by EdgeRyders. The possibility of mixing online and offline (on-site workshops) conversations and activities has also been considered. At the

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688670.

moment of writing, the available data only concerned online conversations occurring on https://edgeryders.eu/en/op3ncare/home

In all cases, the data describes how people either propose ideas (launch a thread) or build from other participants' comments[§].

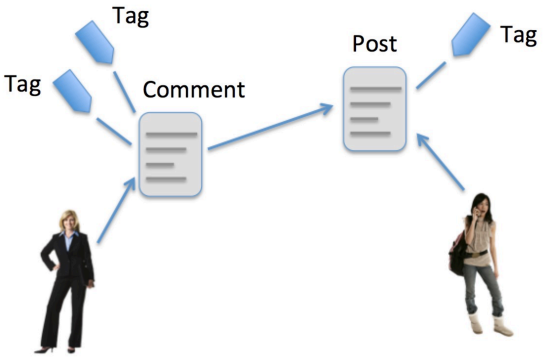

Figure 1. The data records collective discussion dynamics.

As far as the EdgeRyders portal is concerned, a Drupal database records all interactions taking place in conversations, as illustrated in Fig. 1.

- Posts or comments are time-stamped;
- Comments relate to the comment and/or post they point to;
- Users author comments and/or posts;
- Content (posts, comments) are moreover tagged with ethnographic codes (sometimes also called tags).
- Tags are associated with annotations, that is, portion of texts selected by ethnographers when tagging content.
- Tags are inserted through annotations (not shown in Fig. 1). An annotation is formed by a portion of text selected in a post or comment to which a tag is associated.

Data abstraction

Having this material in hand, we now must translate high-level questions into more operational actions conducted on the data.

The associations between the various data elements suggest using a central abstraction to explore the data and answer questions: *networks*. A first network structure we may consider is that depicted in Fig. 1, where links embody associations between entities.

A discussion is triggered by an original post (Fig. 2 bottom), and then develops into a tree of comments (Fig. 2 going upwards).

[§] See also our data management plan.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688670.

Users author comments in the thread initiated by a post — a same user can author numerous comments in the thread (including the author of the initial post). Tags are attached to posts and comments. The resulting network structure becomes quite intricate, even for a single thread (see Fig. 4 and Fig. 5).

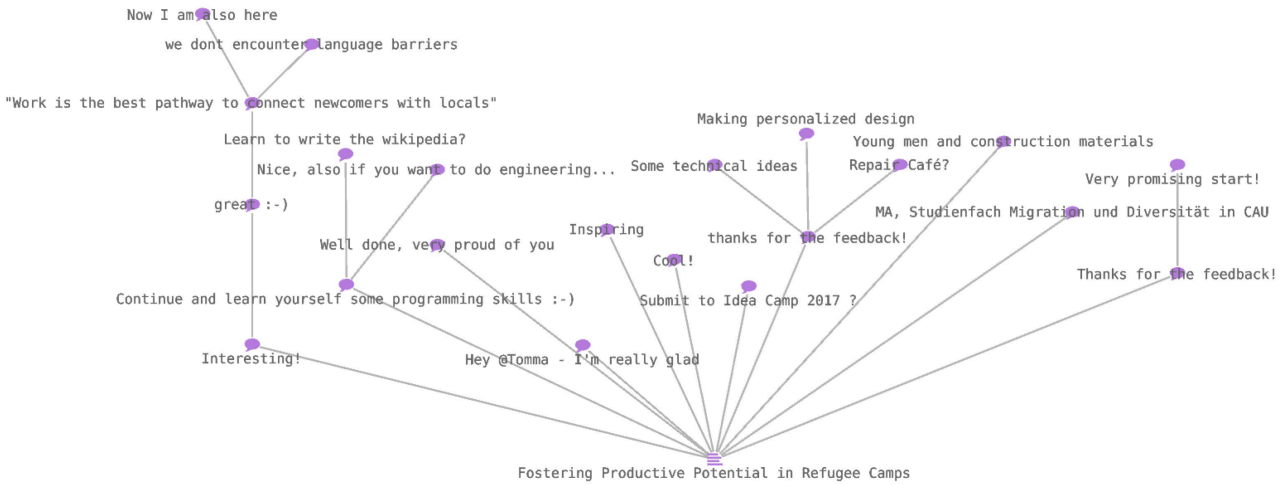

Figure 2. A discussion is triggered by an original post (bottom), and then develops into a tree of comments (going upwards).

The data we deal with gathers all threads, resulting in a "forest" of trees formed by posts and their sprout comments (see Fig. 3).

Figure 3. Threads group into a "forest" of trees formed by posts and their sprout comments.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688670.

These forest become linked to one another since users take part in multiple discussions. Tags worsen the entanglement of the network as posts and comments may share multiple tags (see Fig. 4).

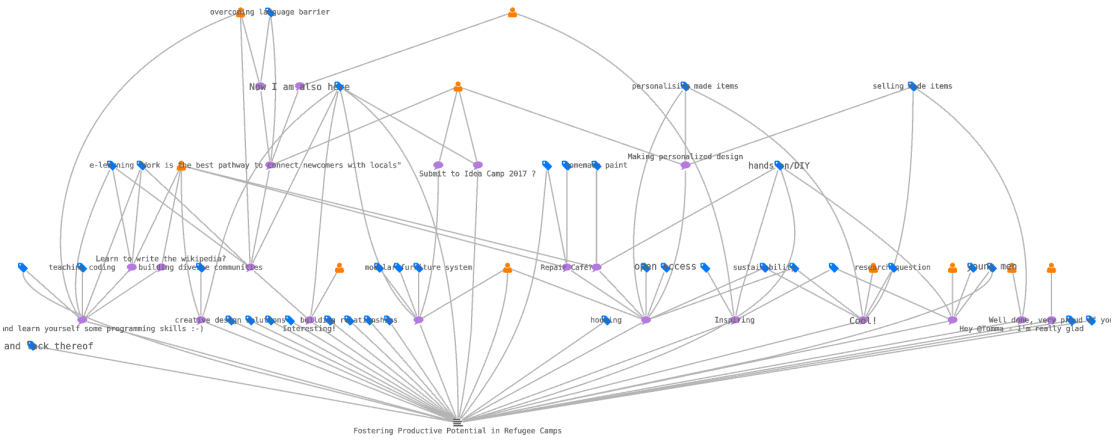

Figure 4. The tree of comments (from Fig. 2) enriched with involved users (orange) and associated tags (blue).

We will refer to this dataset as the *Forum Network* gathering all available information on discussion threads.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688670.

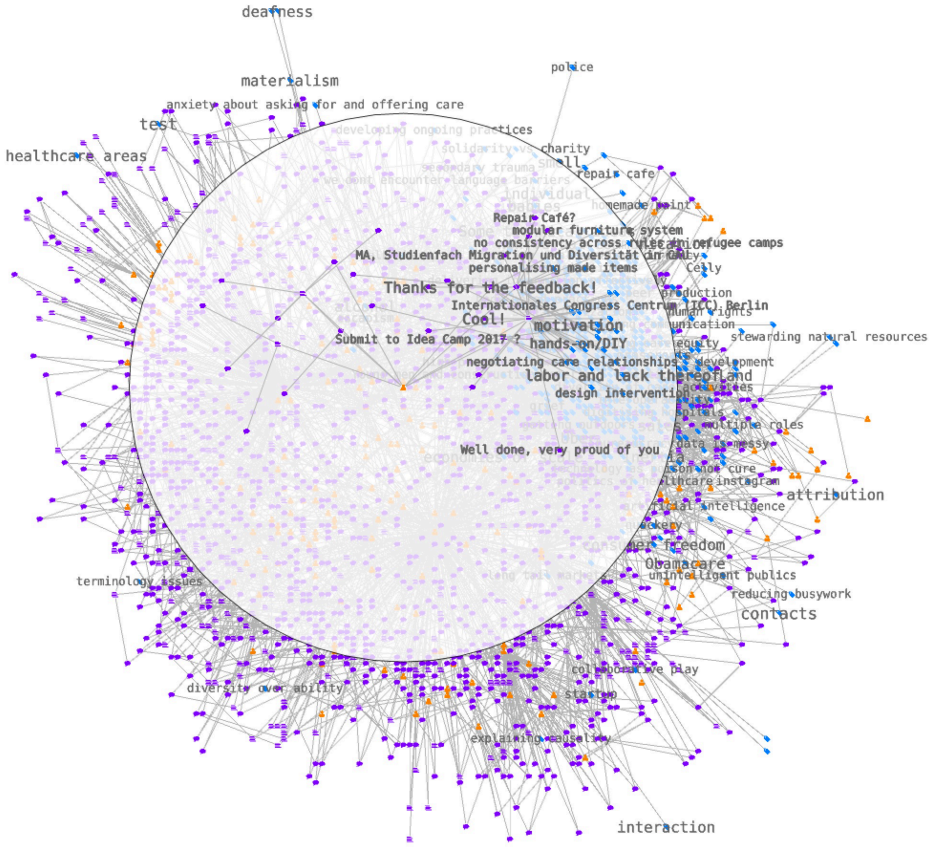

Figure 5. The *Forum Network* gathering all entities posts, comments, users and tags.

The database underlying the Drupal portal hosting the conversations actually stores all of these information. That being said, it also contains lots of information, which does not apply to the analysis we are conducting here (user status and permissions, group structure, etc.).

Also, a relational database is less convenient and less efficient when performing network queries (Angles 2012) (Batra & Tyagi 2012). We consequently decided to derive the network data into an independent graph database using Neo4j (Webber 2012).

Data operations

Let us look at tasks listed for ethnographers — of which many are also relevant to community managers and a larger audience, as mentioned earlier.

Task set #1 are translated into operations that need to be conducted on the network structure.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688670.

In some cases, auxiliary networks need to be computed. One good example is the social network derived from authored content**. Authors A and B get connected A -> B when author A comment content authored by author B. We simply call it the *Social Network*.

We also consider a network connecting tags with one another to support a topic-based navigation of the data. We call this network the *tag co-occurrence network*. The relevance and potential uses of the tag co-occurrence network became quite clear after the Milano workshop (see deliverable D1.2).

This approach of designing and combining different networks derived from the same original data is borrowed from the work of (Renoust 2013) (Renoust *et al.* 2013) and (Renoust *et al.* 2014) on multiplex networks.

Questions / Tasks	Operations and data abstraction
See how tags cover the conversations being examined.	**Use a full or partial view of the forum network.**
Find the most important conversations associated with a set of tags.	**Synchronized views involving a view on (ranked) tags, and the possibility of retrieving the associated discussion threads (filtered or ranked according to importance).**
Find "rich" conversations; find the most "insightful" post, the one engaging the most people.	**Retrieve discussion threads based on** • **Richness index computed on tags** • **Number of people involved in discussion**
Picture the tags most often associated with a person.	**Synchronized views involving a view on (ranked) persons, and the possibility of retrieving a associated tag-tag sub-network, or relevant discussion threads (filtered or ranked according to importance).**
Provide feedback on users "level of expertise".	**Need to design a specific index measuring a user's level of expertise (in terms of topic coverage, number of posts, content length, etc.** **Map this index onto relevant views using well chosen visual variables (see next subsection).**
Tags associated to a given user could be pictured using a tag cloud.	**Can be seen as a companion view to different tasks already mentioned.**

It is this set of different networks, and more importantly the capability of deriving on-demand networks from the original data gathering users and content that we call a *semantic social network*.

** See also Data Management Plan.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688670.

The original network data is thus used to derive a series of different network structures, each capable or more suited to answer different questions and support different tasks.

Task set #2 is similarly addressed. Most questions of this task set correspond to statistics to be computed on the data and/or network.

Questions / Tasks	Operations and data abstraction
Distinguish "rich" posts or comments, those having a larger number of associated tags, and presumably being longer posts.	This requires computing the number of tags part of the neighborhood of a post/comment node in the Forum Network. Length of post/comment refers back to the original data.
The notion of a "popular" tag (associated with more content) also came up as being of interest.	Conversely, we can compute the *degree* (number of neighbors) of a tag node in the Forum Network.
The number of persons involved in a post is an interesting statistics.	This requires retrieving all items involved in a thread (as in Fig. 2); and then compute the number of user nodes it comprises.

As previously mentioned, these lists of tasks are subject to be extended and adjusted as our interactions with users evolve.

Visual Encoding and Interaction Design

The representation of networks as node-link diagrams came as an obvious choice. This type of representation seems relevant when dealing with lower edge density networks (Ghoniem *et al.* 2005) and in some case shows superior user efficiency (Holten *et al.* 2011). Our users (community managers and ethnographers) moreover are accustomed to this type of representation. It is true that early sessions (even before the project started) already made use of these representations since they are central to the tool we used[††].

That being said, showing the whole network at once (as in Fig. 5 above) on the screen may be confusing for the user and inefficient. We thus design different filters and/or dynamic queries that can be applied to the network, resulting in partial, smaller and more readable view of the data.

[††] The *Tulip* Graph Visualization Framework, see http://tulip.labri.fr

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688670.

Forum Network view

The Forum Network itself while being too large to be displayed at once is not that dense. As mentioned in the previous section, it will presumably be used to query discussion threads and show views related to specific users and/or tags. We may also run a level-of-details type of a view on the network by selecting any focus node (user and/or tag) and expand away from the node by selecting neighbors of interest. Inspired by the work of (van Ham & Perer 2009), we designed and implemented early version of this filter (Fig. 6) and will keep improving according to our users' experience and requests.

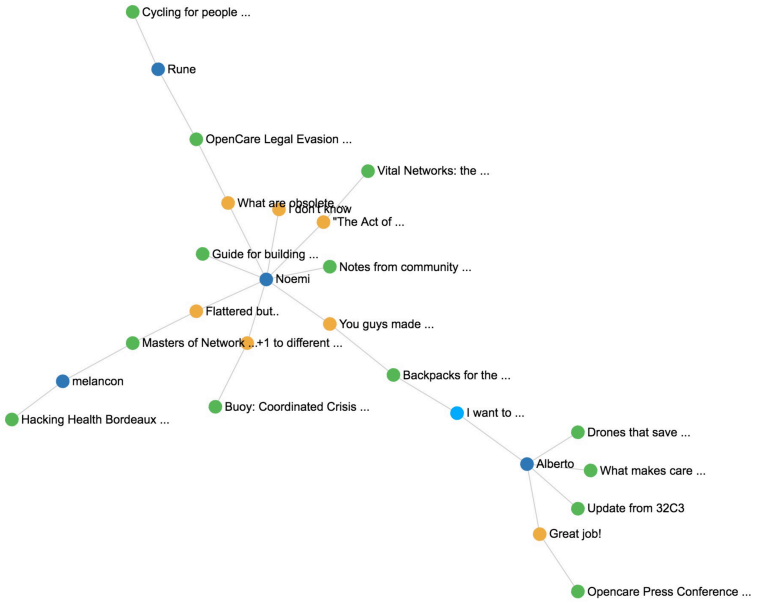

Figure 6. Partial level-of-details view of the Forum Network showing users (blue), posts (green) and comments (soft yellow). See http://164.132.58.138:9000/#/dashboard/doi

For now, users can specify any term from which the view is built (user name, content title of keyword). The sub-network is then computed according to a relevance metric mixing structural information (node degree or centrality, for instance) and data attributes. User feedback will be crucial to relevantly design the metric guiding the computation of this level-of-detail view.

Tag Co-occurence view

After the Hacking Health Bordeaux sprint, and certainly after the Milano consortium meeting, users confirmed the tag co-occurrence as a useful device. For community managers and more generally for users: to reveal the dynamics of the conversation, showing how issues mix in online debates; for ethnographers, to provide feedback on their own work.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688670.

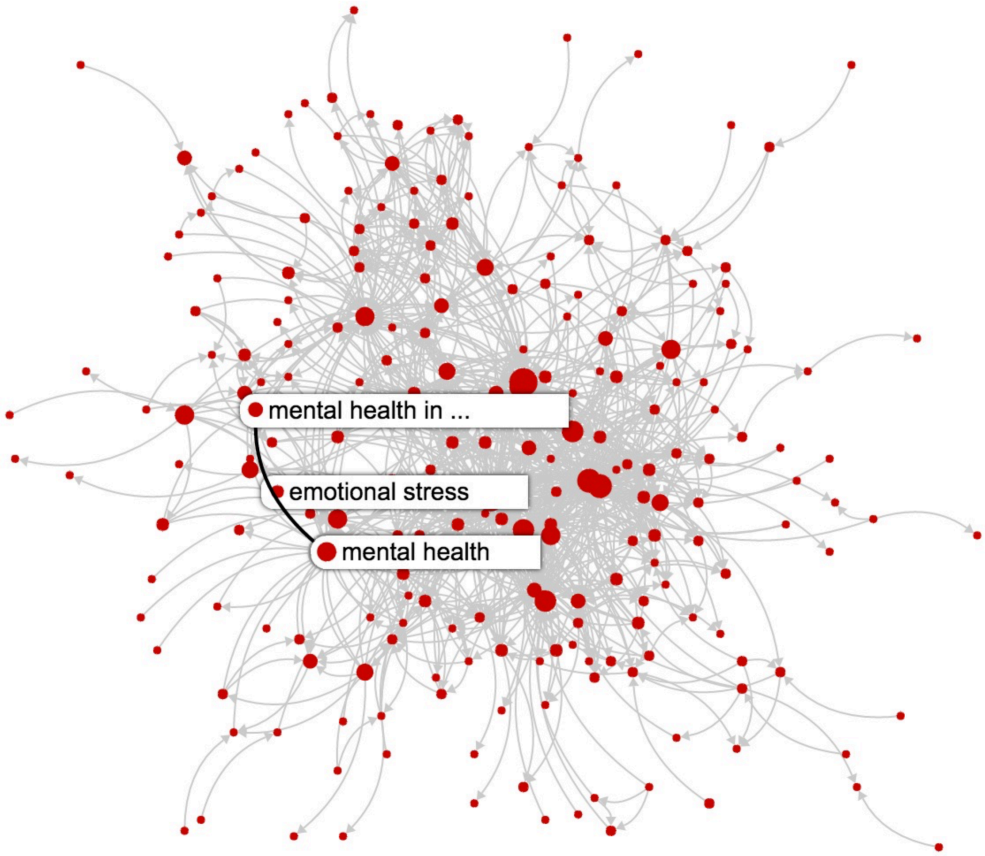

Figure 7. Tag co-occurence network.

The tag co-occurrence network can potentially be quite dense, as any pair of tag may connect. For sake of comparison, the Forum Network comprises about twice more edges than nodes, while this ratio exceeds 12 or 13 for the tag co-occurrence network!

With such a high edge density, any layout approach is hardly capable of producing an optimally readable map of the network (see Fig. 7) as the abundance of edges turns the layout computation into an over-constrained optimization problem. One approach is to filter out edges and deal with a lower density sub-network. Approaches such as that of (Nick *et al*. 2013) are now being tested. We also plan to implement and validate level-of-detail approaches on the tag co-occurrence network.

Time filter and other statistics

As we have mentioned, since content is time-stamped, so are the various networks we derive from the original data. We may thus draw different graphs showing how networks statistics evolve over time (number of posts/comments, or any structural statistics on the network).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688670.

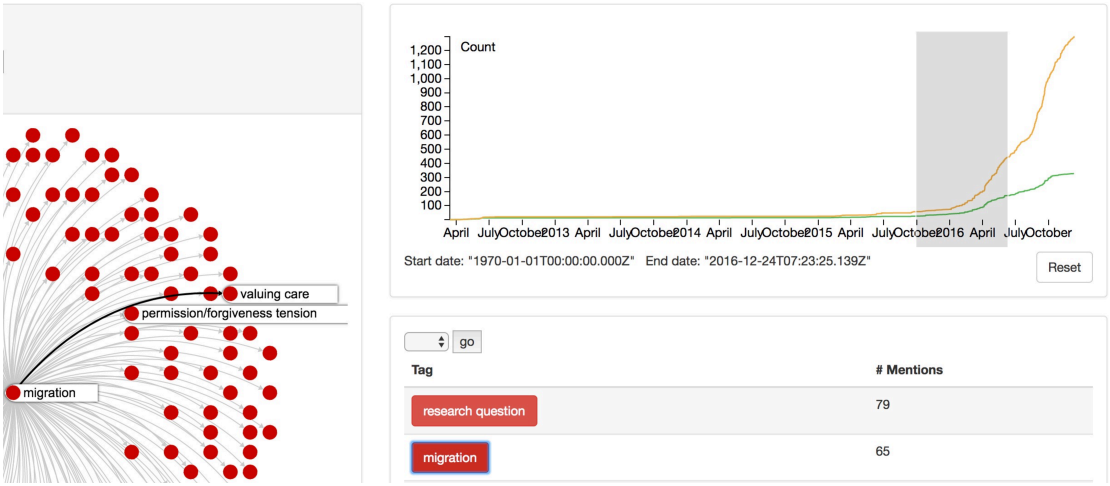

Figure 8. A timeline is integrated in all views (when applicable). Views are filtered according to a selected time interval, and can be inspected by sliding the time window.

Alternate views can also help compare statistics with one another (see http://164.132.58.138:9000/#/dashboard/globalView).

Interaction

Several GUI interactions have been experimented, starting with the basic pan & zoom. Being able to closely inspect the neighborhood of a node is central (understanding why some nodes are connected or close). Computing a partial level-of-detail is one way of addressing this requirement. Another is to provide a customized lens, bringing neighbors upfront while pushing other nodes in the background (see Fig. 5 above).

Although basic, selecting, hovering or clicking on entities to access the underlying data is quite crucial.

Synchronized views supports querying a view by means of a secondary display of data such as node ranking. An example is selecting tags from a ranked list and displaying "rich" threads involving these tags. Another example is selecting a group of close tags and displaying the social sub-network of people interacting around these tags. Conversely, the user could select a group of close users and tags associated with threads involving them could be displayed using a tag could.

The choice of the most efficient interactions for each task will follow from user experiment. Interactions that may please community managers may not be suitable for ethnographers or a larger user audience.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688670.

Visual variables

Selecting the best possible layout to position entities on the screen is crucial — graphical entity positions ranks first among all visual variables (Mackinlay 1986).

The display of network statistics on the side with synchronized view can alternatively be addressed by mapping statistics or data attributes directly on network entities. Visual variables such as node size and hue, or gradient is a classical approach to map numerical/ordinal/categorical information. Edge thickness has often been mentioned by users to reflect the intensity of connections (number of co-occurences between tags, or number of interactions between users).

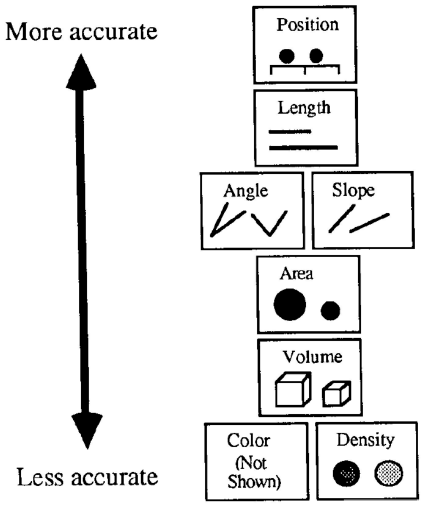

Figure 9. MacKinlay's accuracy ranking of quantitative perceptual tasks.

Finding the right trade-off between all these constraint (position, colormaps, length/thickness, etc.) — starting with guidelines and a set of best practices (Mackinlay 1986) (Harrower & Brewer 2013) — also needs to be experimented by users before we can claim they are the right choice.

Conclusion

Using Munzner's approach proves to be quite efficient in that it brings us to design and develop more effective visualizations. Many of the visualizations have been first tested through prototypes in live sessions and will now be integrated into a web-based environment (the dashboard or the OpenEthnographer companion). A preliminary view of the dashboard is available here: http://164.132.58.138:9000/

At the time of writing this document, many of the tasks described herein were designed, implemented and tested by users. We use Munzner's four-level model in an iterative manner, designing views and interaction incrementally, trying to validate each design choice with users before engaging in further development.

In doing so, task requirements are adjusted and/or refined/expanded, newt tasks may emerge after users see the potential they are offered with the SSNA dashboard.

References

Angles, R. (2012). A Comparison of Current Graph Database Models. Data Engineering Workshops (ICDEW), 2012 IEEE 28th International Conference on, 171-177.

Batra, S. and C. Tyagi (2012). "Comparative Analysis of Relational And Graph Databases." International Journal of Soft Computing and Engineering (IJSCE) 2(2): 2231-2307.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688670.

Card, S. K., J. D. Mackinlay, B. Shneiderman. (1999). Readings in Information Visualization. San Francisco, Morgan Kaufmann Publishers.

Cottica, A., G. Melançon, et al. (2017). Testing for the signature of policy in online communities. Complex Networks & Their Applications V (Cherifi *et al.* eds.), Studies in Computational Intelligence, volume 693. Springer, pp. 41-54.

Ghoniem, M., J.-D. Fekete, et al. (2005). "On the readability of graphs using node-link and matrix-based representations: a controlled experiment and statistical analysis." Information Visualization (Palgrave) **4**(2): 114-135.

Harrower, M., & Brewer, C. A. (2013). ColorBrewer. org: an online tool for selecting colour schemes for maps. The Cartographic Journal.

Herman, I., M. S. Marshall, G. Melançon. (2000). "Graph Visualisation and Navigation in Information Visualisation: A Survey." IEEE Transactions on Visualization and Computer Graphics 6(1): 24-43.

Holten, Danny, et al. (2011) "An extended evaluation of the readability of tapered, animated, and textured directed-edge representations in node-link graphs." 2011 IEEE Pacific Visualization Symposium, pp. 195-202.

Mackinlay, J. (1986). "Automating the design of graphical presentations of relational information." Acm Transactions On Graphics (Tog) 5(2): 110-141.

Meyer, M., M. Sedlmair, T. Munzner. (2012). The four-level nested model revisited: blocks and guidelines. Proceedings of the 2012 BELIV Workshop: Beyond Time and Errors - Novel Evaluation Methods for Visualization. Seattle, Washington, ACM: 1-6.

Munzner, T. (2009). "A Nested Process Model for Visualization Design and Validation." IEEE Transactions on Visualization and Computer Graphics 15: 921-928.

Nick, B., C. Lee, et al. (2013). Simmelian Backbones: Amplifying Hidden Homophily in Facebook Networks. Advances in Social Network Analysis and Mining (ASONAM 2013), pp. 525-532.

Renoust, B. (2013). Analysis and Visualisation of Edge Entanglement in Multiplex Networks, Doctoral dissertation, University of Bordeaux.

Renoust, B., Melançon, G., Viaud, M. L. (2013). Measuring group cohesion in document collections. In IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), vol. 1, pp. 373-380.

Renoust, B., G. Melançon, Viaud, M. L. (2014). Entanglement in Multiplex Networks: Understanding Group Cohesion in Homophily Networks. Social Network Analysis - Community Detection and Evolution. R. Missaoui and I. Sarr, Springer International Publishing: 89-117.

Sedlmair, M., M. Meyer, T. Munzner. (2012). "Design Study Methodology: Reflections from the Trenches and the Stacks." IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis 2012) 18(12): 2431-2440.

van Ham, F. and A. Perer (2009). "Search, Show Context, Expand on Demand: Supporting Large Graph Exploration with Degree-of-Interest." IEEE Transactions on Visualization and Computer Graphics 15(6): 953-960.

Wehrend, S. and C. Lewis (1990). A problem-oriented classification of visualization techniques. 1st conference on Visualization '90, San Francisco, California, IEEE Computer Society Press, 139-143.

Webber, J. (2012). A programmatic introduction to Neo4j. Proceedings of the 3rd annual conference on Systems, programming, and applications: software for humanity, Tucson, Arizona, USA, ACM, 217-218.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688670.