The Importance of Including Collective Intelligence in the Open Data Management Lifecycle

View point by Cristina Sarasua

Switzerland's extensive investment in Open Data Initiatives has led to a rich landscape with thousands of open data sets[1] in domains ranging from urban statistical data to mobility planning and cultural information. Anyone willing to develop an app or run a data analysis can browse, download and query the (linked) open data via the data portals and technical infrastructure (e.g., query endpoints and APIs) provided by public organisations such as the statistical offices of the city[2] and the canton of Zurich[3].

With the exception of some data providers who tend to publish updates of their data, the data preparation and publishing process is often:

- *Non-iterative*: the data is published and it is not further corrected, refactored or conceptually extended.
- *Unilateral*: the only agent involved in structuring, cleaning, opening and documenting the data is the data provider, who despite being knowledgeable in the data, cannot cover all use cases and only single user's/user group's knowledge (or point of view) can lead to data bias (Baeza-Yates, 2018).
- Relying on purely *automatic methods*: when data publishers need to e.g., link their data to other data sets they use automatic tools because these tools help them save time.

While this process results in already valuable data sets, a more dynamic and hybrid process, interleaving human and machine intelligence would aid in improving not only data quality, but also data diversity. Machines are good at imitating human assessment (e.g., in machine learning), executing search efficiently and running routines repeatedly, but humans are still more accurate at certain tasks such as performing complex associations of ideas, identifying quality, evaluating relevance and being creative. Hence, if we *involve many and diverse people in the open data management* process, they can find and correct errors and data holes iteratively, they can identify needs for data extension or data re-annotation, they can connect the data in various ways, they can suggest different interpretations of the data and they can find new ways to use the data. **We need to fuel the open data management lifecycle with *collective intelligence*** --- a term defined by Malone and colleagues as "groups of individuals doing things collectively that seem intelligent" (Malone et al., 2010), that is closely related crowdsourcing (i.e., with individuals of the general public who are involved via an open call) and human computation (i.e., humans solving problems that machines cannot solve) (Quinn and Bederson, 2011).

Designing and running a collective intelligence-based system is not as straightforward as implementing some code, scheduling its execution and waiting for its output, though. It is way more difficult, especially since in collective intelligence (and by extension in crowdsourcing) systems we need to handle humans' motivational, cognitive and error diversity (Bernstein, 2012). To achieve the full potential of collective intelligence, one needs to take care of crowd quality control (Daniel et al., 2018), develop participant retention and engagement mechanisms, find the right combination of

[1] https://opendata.swiss/en/
[2] https://data.stadt-zuerich.ch/
[3] https://statistik.zh.ch/internet/justiz_inneres/statistik/de/daten/opendata.html

people, stimulate healthy social dynamics, and understand how to help people become as effective and efficient as possible, as well as smartly design how to combine human and machine computation. There are many successful examples that one can learn from. For instance, CrowdLang, the system that managed to translate one book in one hour reaching an accuracy close to professional (Minder and Bernstein, 2012), and Zooniverse, the project in which amateurs in astronomy analyse pictures taken by telescopes and discovered a new cluster of galaxies (Lintott et al., 2008). Another reference example is Wikidata[4], the free, crowdsourced and multilingual knowledge base that anyone in the world can query and edit. Wikidata is a project where thousands of volunteers have collaboratively curated and maintained a data set with more than 49 data items in five and a half years, and where bots are implemented and supervised by them (Vrandečić and Krötzsch, 2014).

Who should these people be?

Research has shown, both in academic and industrial contexts, that it is possible to involve non-professional crowd in data management tasks via paid and volunteer crowdsourcing (Marcus and Parameswaran, 2015) (Sarasua et al., 2015) (Demartini et al., 2017). However, if there is the opportunity to simultaneously involve the geographer who can help run a deep spatial analysis, the statistician who can design a detailed exploratory data analysis, or the librariarian who is proficient in cataloguing, it would be a mistake not to involve them, because the ideal way to design collective intelligence is to combine crowdsourcing with experts' help, either for guiding the process, training the non-experts, or filtering / ranking / organizing crowd actions.

Indeed, Switzerland is in a privileged position, as it hosts plenty of hackathons (e.g., MakeZurich[5], Wikidata Zurich Datathon[6] and Hackathon[7], TWIST[8], Hack'n'Lead[9], ODD Hackathon[10]), meetups, initiatives (e.g., Open Data CH[11], School of Data[12], Open Data Zurich[13], OGD Canton Zurich[14]) and technical conferences (e.g., DINAcon[15]) where skilled developers and open data enthusiasts design, implement and discuss technology and data. These people already contribute in independent (and sometimes inter-connected) projects. Still, this activity does not guarantee the iterative, hybrid and collective data management process that can increase the value of the open data sets that have already been published. So, let's introduce the means to facilitate these people's (and others') collective action towards an open data management process in constant evolution --- these being technology, new methodologies and social events.

Proposals for Open Data CH Enthusiasts

We need to (i) design orchestration and task-oriented cooperation, (ii) enable and manage collective intelligence for this open data management scenario and (iii) monitor the progress that we achieve collectively with regard to evolving, improving, and analysing the open data.

I propose to work on the following actions, as they could help us make a step forward in this direction:

[4] https://www.wikidata.org/
[5] https://makezurich.ch/
[6] https://www.wikidata.org/wiki/Wikidata:Events/Wikidata_Zurich_Datathon
[7] https://www.wikidata.org/wiki/Wikidata:Events/Wikidata_Zurich_Hackathon
[8] http://twist2018.ch/
[9] http://hackandlead.com/
[10] https://zurich-r-user-group.github.io/hackathon.html
[11] https://opendata.ch/
[12] https://schoolofdata-ch.github.io/
[13] https://data.stadt-zuerich.ch/
[14] https://statistik.zh.ch/internet/justiz_inneres/statistik/de/daten/opendata.html
[15] https://dinacon.ch/en/

- Extend Data Portals to Enable Structured Discussions (that can be queried a posteriori), to collect ideas and feedback that help identify (i) new data that can complement or extend current data sets, (ii) issues that occur when using or analysing the data sets hosted in the data portal. For both things, proposing discussion seeds and having a human (or semi-automatic) moderation method will be key. For such discussions to be productive, there should be a method that maximizes *collective creativity*.
- Promote and organise the continuation of data analysis on data sets: discovering a fact about a data set is useful, but we should generate conversations between data analyses. Not only do different people analyse data differently (Feldman et al., 2018), but also different people might have different motivations to analyse the data. Moreover, it is of utmost importance that the findings, the data and the analysis are linked. The work by Garijo and Gil (Garijo et al., 2017; Belhajjame et al., 2015) provides many ontologies to annotate this information as structured data..
- Do not expect everyone to be actively participating. Empirical research shows that frequently in crowdsourcing systems few people do most of the work, and those who become power contributors have a constant commitment (Sarasua et al., 2018; Panciera et al., 2009), probably because they are intrinsically motivated. Therefore, it is important to target people who are committed to the values of open knowledge, convince those who are not yet convinced, and try to make the behaviour of power users contagious.
- Design social systems that encourage a productive interaction between the members of the group. For example, a study by Woolley and colleagues found that the average social sensitivity of group members is related to the group's collective intelligence factor, which is "the general ability of the group to perform a wide variety of tasks" (Woolley et al., 2010). When it comes to innovation and ideation processes, the ethnographer Linda Hill[16] highlights the importance of enabling a space for constructive criticism to boost collective creativity, as diverse people can come up with many different alternative ideas.

I believe that many more proposals are about to come. We will need to develop new technology to make these ideas happen. Hopefully they will be defined *collectively*.

There will be a unique upcoming opportunity to learn about collective intelligence, crowdsourcing and human computation, as two interdisciplinary and co-located conferences on the topic will take place in Zurich:

- ACM Collective Intelligence Conference (CI) 2018[17], taking place on July 7-8 at the University of Zurich.
- Human Computation and Crowdsourcing (HCOMP) 2018[18], taking place on July 5-8 at the University of Zurich.

[16] https://www.ted.com/talks/linda_hill_how_to_manage_for_collective_creativity

[17] https://ci.acm.org/2018/

[18] https://www.humancomputation.com/2018/

Ricardo Baeza-Yates. 2018. Bias on the web. Commun. ACM 61, 6 (May 2018), 54-61. DOI: https://doi.org/10.1145/3209581

Malone, T. W., Laubacher, R., & Dellarocas, C. (2010). The collective intelligence genome. *MIT Sloan Management Review*, *51*(3), 21.

Quinn, A. J., & Bederson, B. B. (2011, May). Human computation: a survey and taxonomy of a growing field. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1403-1412). ACM.

Adam Marcus and Aditya Parameswaran. 2015. Crowdsourced Data Management: Industry and Academic Perspectives. Found. Trends databases 6, 1-2 (December 2015), 1-161. DOI=http://dx.doi.org/10.1561/1900000044

Sarasua, C., Simperl, E., Noy, N. F., Bernstein, A. & Leimeister, J. M. (2015). Crowdsourcing and the Semantic Web: A Research Manifesto. Human Computation Journal.

Demartini, G., Difallah, D. E., Gadiraju, U. & Catasta, M. (2017). An Introduction to Hybrid Human-Machine Information Systems.. *Foundations and Trends in Web Science*, 7, 1-87.

Bernstein, A., Klein, M. & Malone, T. W. (2012). Programming the global brain.. Commun. ACM, 55, 41-43.

Lintott, C. J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., ... & Murray, P. (2008). Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, *389*(3), 1179-1189.

Minder, P., & Bernstein, A. (2012). How to translate a book within an hour: towards general purpose programmable human computers with CrowdLang. In Proceedings of the 4th Annual ACM Web Science Conference (pp. 209-212). ACM.

Vrandečić, D., & Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, *57*(10), 78-85.

Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. 2018. Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques, and Assurance Actions. ACM Comput. Surv. 51, 1, Article 7 (January 2018), 40 pages. DOI: https://doi.org/10.1145/3148148

Feldman, M., Anastasiu, C. & Bernstein, A. (2018). Towards Collaborative Data Analysis with Diverse Crowds - A Design Science Approach.. In S. Chatterjee, K. Dutta & R. P. Sundarraj (eds.), *DESRIST* (p./pp. 218-235), : Springer. ISBN: 978-3-319-91800-6

Belhajjame, K., Zhao, J., Garijo, D., Gamble, M., Hettne, K. M., Palma, R., Mina, E., Corcho, Ó., Gómez-Pérez, J. M., Bechhofer, S., Klyne, G. & Goble, C. A. (2015). Using a suite of ontologies for preserving workflow-centric research objects.. *J. Web Sem.*, 32, 16-42.

Garijo, D., Gil, Y. & Corcho, Ó. (2017). Abstract, link, publish, exploit: An end to end framework for workflow sharing.. *Future Generation Comp. Syst.*, 75, 271-283.

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. science, 330(6004), 686-688.

Katherine Panciera, Aaron Halfaker, and Loren Terveen. 2009. Wikipedians are born, not made: a study of power editors on Wikipedia. In Proceedings of the ACM 2009 international conference on Supporting group work (GROUP '09). ACM, New York, NY, USA, 51-60. DOI=http://dx.doi.org/10.1145/1531674.1531682

Cristina Sarasua, Alessandro Checco, Gianluca Demartini, Djellel Difallah, Michael Feldman and Lydia Pintscher. (2018). The Evolution of Power and Standard Wikidata Editors: Comparing Editing Behavior over Time to Predict Lifespan and Volume of Edits. To appear in: Journal of Computer Supported Cooperative Work, Springer.