Coordination of Biological and Chemical IT Research Activities

FP7-ICT-2009.8.9 Coordinating Communities, Plans and Actions in FET Proactive Initiatives
Project No. 270371

www.cobra-project.eu

Collective Intelligence Site

WP1 D1.3

March 23 2012

Workpackage	1
Due Date	M15
Date of Release	23/3/12
Deliverable Type	O
Deliverable Number	1.3
Pages	6
Leading Unit	SDU
Leading Author	Steen Rasmussen
Contributing Authors	Mark Dorr, Sif Schmidt-Petersen, Harold Fellemann and John McCaskill
Contributing Units	RUB
Contact Mail	steen@sdu.dk

COBRA is a Future and Emerging Technologies Proactive Initiative funded by the European Commission under FP7.

Method for mapping the collective intelligence of the ChemBio-ICT stakeholders and the detailed design infrastructure of the online survey

Lead Institution: SDU
Contributing institution: RUB
Lead Author: Steen Rasmussen, steen@sdu.dk
Contributing authors: Mark Dorr, Sif Schmidt-Petersen, Harold Fellemann & John McCaskill

1. Method

We have developed an information and communication technology (ICT) based collective intelligence method, which stems from several different traditions including the facilitation of citizen participation, the study of social groups, the use and development of survey methods, automated language parsing as well as investigation of the many new web 2.0 possibilities. An earlier version of this web-based method was used to gage the collective intelligence of the Artificial Life community in connection to the Artificial Life VII Conference in Portland, summer 2000 (Rasmussen et al., 2003). This method has also been used to for strategic research planning for the Earth and Environmental Division at Los Alamos National Laboratory NM USA (Keating et al., 2001a), for Governance Efficiency Studies in the Navajo Nation (Keating, Rasmussen & Raven, 2001) as well as for assessing the best way to build up Geographic Information Systems (GIS) based decision support for Los Alamos National Laboratory.

The method works through a series of steps, as illustrated in the figure below. The method is able to map the "lay of the land" for any complex set of issues within a large stakeholder community. The method is fast and inexpensive as all input comes from an online survey interface where the stakeholder community defines their own issues.

1. A small, diverse, and representative subset of the stakeholder group designs an initial information repository, including formulating key questions about the problem complex on the Web. Next all stakeholders in the larger community individually review the information about the issue complex, either through the associated Web environment or through town hall meetings, media, conversations, and so on. Where possible, stakeholders add information about the problem context to the Web storehouse for review by others.

2. All stakeholders in the group rank and organize issues relevant to the problem and express their opinions about the issues through an online, open-response survey that allows freely typed input to questions. Individuals can describe new issues as well as rank possible already defined issues.

3. The feedback from step 2 is parsed, synthesized and analyzed to identify possible areas of conflict and consensus via graphical frequencies, a variety of statistical correlations, mind maps, and other relevant plots. This analysis can increasingly be done automatically online by using an off the shelf language parser to extract the pertinent concepts from the open text input provided by the stakeholders. These concepts can then automatically go through statistical analysis and eventually be graphed.

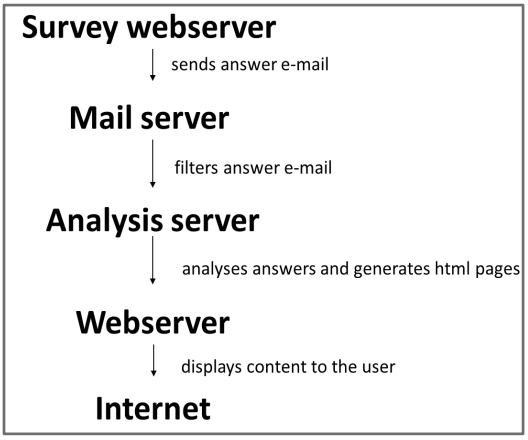
4. The results of the analysis, the graphs and statistical findings, are made available to the stakeholders at large through the Web.

5. The condensed collective intelligence now gathered in the data repository, through the above-described process, allows the whole community to make decisions based on the information resulting from the analysis of all the answers from the survey. The process is transparent and bottom up.

Steps 2–4 can be repeated as the group reacts to areas of conflict and agreement, and as individuals modify their positions. Once a group has clarified its conflicts and identified its areas of consensus, it can take action on these matters. It is our contention that this sort of self-organizing collective intelligence process enables a group to make better-informed decisions about the important problem complexes that it faces.

2. Design details for the online survey

A special aspect of the collective-intelligence process described above is the open-ended responses allowed in step 2. The open-response survey does three important things: It organizes stakeholder input along a set of broadly defined questions about the set of issues; it allows open-ended input; and it limits each response of each individual to a few sentences. Once the open-response data has been gathered, it is brought into a quantifiable form by coding each response into one of a finite number of response categories either done manually or via statistical methods. We use the python programming language together with the open source library called the Natural Language Toolkit (NLTK, http://www.nltk.org), see Bird, Klein and Loper, 2009, to parse the text and identify the key concepts or concept combinations. The NLTK is an easy-to-use, well documented concept, and its in-line script processing is well in sync with the workflow of the server infrastructure system. When doing computational linguistics, or natural language parsing, there are a number of ways to process a body of text. The NLTK allows us to tokenize a text, i.e. splitting it into correct grammatical categories; stemnize, that is, reducing verbs to their stem-form; filter the text and ask for specific categories of words; and by use of the semantic module it also makes it possible to extract meaning out of a text.



The automated data analysis flow and server infrastructure, as shown in the figure to the left, consist of five parts: (i) survey webserver for question presentation, with user management (tracks user activities), (ii) mail server for message forwarding, sorting and backup, (iii) analysis server for running analysis software in a secured environment (bash, python (jinja2, matplotlib, nltk), c++, java,...), (iv) webserver for displaying the html content, (v) internet for distributing the content. The security of each step of the process is as follows: (i) The survey webserver is protected by the company Rambøl. (ii) The mail server is protected by the firewall of the University of Southern Denmark (SDU). (iii) The analysis server is not connected to the Internet and is located at the FLinT center at SDU. (iv) The webserver for displaying the analysis content is protected by SDU and only allows public viewing.

The developed automated survey analysis capability means that the data analysis of the input from the current stakeholder can be made available for the stakeholder immediately after the completion of the survey. The stakeholder will thus directly be able to review his/hers input after the survey is completed together with the previous inputs into the survey database.

The automated survey analysis feedback of course operates both with predefined questions and with open-response questions, see Appendix. To make the ChemBio-ICT survey as robust as possible we plan to use both open-response and predefined questions and as a backup we can use human review of the open-response questions in case the automated language parser does not operate as desired.

Familiar statistical and other methods (e.g. support vector machines and Bayesian filters) can be employed to extract information from the open-response data once it has been categorized.

The key feature of the open-response survey is the ability to take input that is completely open in content and restricted only in length. The open-response survey can be thought of as a “fishing net” that efficiently and inexpensively catches all the worries, excitement, visions, complaints, and the like in the group and makes them available for both qualitative and quantitative analysis. This mitigates the familiar bias in traditional surveys caused by forcing all responses to be chosen from predefined answers to predefined questions.

References

G. Keating, S. Rasmussen & M. Raven (2001a). Web-based consensus building and conflict clarification for EES division's strategic planning process: New technical directions (Report LA-UR-02-3830). Los Alamos National Laboratory.

G. Keating, S. Rasmussen & M. Raven (2001b). *Consensus building tools for GIS design* (Report LA-13894-MS). Los Alamos National Laboratory.

G. Keating, S. Rasmussen, M. Raven, E. Tso, J. Cocq & P. Dotson (2001). Use of web-based consensus building and conflict clarification process for the Navajo Nation Governmental Eféciency Study, an appendix to *ETD environmental consulting, report of the Navajo Nation Council Evaluation*, submitted to the Oféce of Navajo Government Development, Window Rock, AZ (Report LA-UR-01-6207). Los Alamos National Laboratory.

S. Rasmussen, M. Raven, G. Keating & M. Bedau, Collective intelligence of the Artificial Life community on its own successes, failures and future. Artificial Life 9: 207–235 (2003).

S. Bird, E. Klein & E. Loper, Natural Language Processing with Python, Analyzing Text with the Natural Language Toolkit, O'Reilly Media, June 2009

http://www.nltk.org/

Appendix

Planned questions for the ChemBio-ICT community survey:

Q1 Questions about demographics (all from click menus):

1.1 Age

1.2 Gender

1.3 Geography

1.4 Discipline

1.5 Academia/Industry

1.6 Position

1.7 Webpage

Q2 Questions about professional activities (open-response text)

2.1 List your primary and two secondary areas of expertise

2.2 List references to your three most important papers with relevance to the emerging ChemBio-ICT area (please provide full citation with title and all authors if possible and include doi if available)

2.3 List funded projects you are currently associated with (names and web addresses)

2.4 Check your involvement in domains and their main objectives: based on list of 8 with room for extensions.

2.5 Rate the importance of metrics for ChemBio-ICT and how you would quantify them: based on list of 8 with room for additions.

Q3 Questions regarding your ideas/vision about the emerging ChemBio-ICT community (open-response text)

3.1 Please provide three critical or grand *scientific* challenges that the community is facing now or in the near future.

3.2 Which are, in your view, the three most promising existing and emerging *technologies* from/for ChemBio-ICT? (check box whether "from" or "for").

3.3 Please state the main community and/or organizational stumbling blocks you experience as you seek to make progress in the ChemBio-ICT area. Please address issues specific to the ChemBio-ICT area and not to ICT research and development in general.

3.4 Please list three major medium-short term applications of ChemBio-ICT, and state what breakthrough they depend on, and when you expect it?

3.5 Please list three positive, longer term societal impacts you believe could be developed from ChemBio-ICT.

3. 6 Which potential adverse societal impacts do you believe should be avoided by the emerging ChemBio-ICT activities?

Q4 Questions regarding domains and their objectives.

(This section aims to establish both the objectives and the metrics for "success" for particular domains that we have identified as important or relevant. The Objectives and Metrics tables to be included in the survey have the structure and content domains as shown below, although the final formulations may change.).

Please assign the primary domains of your research interest in chem/bio ICT to one or more of the areas listed on the left. If your objective differs from the overarching one given for this area, please specify. If you wish, please add up to 3 new domains at the same level of generality.

Answering this question will help us to weight and include your domains and objectives in the upcoming roadmap.

Domains and objectives of chem/bio IT

Domain	**Common main objective**	**Your objective, if different**	**Tick if involved in this domain**
Detection & monitoring	Analysis		
Energy transformation & storage	Novel integration		
Informational materials: evolution & learning	Intelligence		
Fabrication & synthesis	Automation		
Material recycling	Conservation		
Sensing, signalling & communication	Coordination		
Control & management	Regulation		
Application development & commercialization	Cost effectiveness		
Other domain 1	Your objective1		
Other domain 2			
Other domain 3			

Q5 Which of the properties below do you think important as metrics for evaluating progress in chem/bio ICT? If there are others, please list up to three that you find most important.

For each of these quantifiable indicators that you deem important, please state how it could be quantified as a metric.

Answering this question will also help us to weight and include your metrics for success in the upcoming roadmap.

Metrics of chem/bio IT.

Domain	**Metric**	**Alternative Metric?**	**How would you quantify this?**	**Tick if involved in this domain**
Detection & monitoring	Universality			
Energy transformation & storage	Efficiency / energy density			
Informational materials evolution & learning	Information density / efficiency			
Fabrication & synthesis	Individual customization			
Material recycling	Sustainability			
Sensing, signalling & communication	Bandwidth / precision			
Control & management	Robustness			
Application development & commercialization	Customer potential			
Other domain 1	Your metric1			
Other domain 2				
Other domain 3				