This paper was published by IEEE-Xplore. Please cite as follows:
Muller, Lars; Wetzel, Thomas; Hobohm, Hans-Christoph; Schrader, Thomas: "Creativity Support Tools for Data Triggered Hypothesis Generation. 2012 Seventh International Conference on Knowledge, Information and Creativity Support Systems (KICSS).
pp. 24-27, 8-10 Nov. 2012.
doi: 10.1109/KICSS.2012.12

keywords: Biomedical imaging; Data models; Data visualization; Problem-solving; Semantics; Technological innovation; biomedical research data; creativity support; data driven science; hypothesis generation

Creativity Support Tools for Data Triggered Hypothesis Generation

Lars Müller
Thomas Wetzel
Hans-Christoph Hobohm
Faculty of Information Sciences
University of Applied Sciences Potsdam
Potsdam, Germany

Thomas Schrader
Department of Informatics and Media
University of Applied Sciences Brandenburg
Brandenburg, Germany

Abstract — **The shift from hypothesis-driven to data-driven science implies new methods for hypothesis generation. Creativity support has to be offered for problem finding based on research data. DataCreativityTools (DCT) is a research & development project building a pilot creativity support tool for the OpEN.SC biomedical data**.

data driven science; creativity support; biomedical research data; hypothesis generation

I. INTRODUCTION

Some years ago Kell and Oliver asked in an essay: "Here is the evidence, now what is the hypothesis?" [1] Large data aggregations allow the shift from hypothesis-driven to data-driven research. This change, though, has not been carried out yet satisfyingly. Scientists need to explore new paths to benefit from these data. We believe that findings from research on fostering creativity will enable us to present research data within a creativity enhancing environment, thus supporting research data triggered hypothesis generation.

II. BIOMEDICAL RESEARCH DATA

Research and development in medical sciences depend increasingly on access to research data. [2] Data comes from various clinical departments and especially from pathology due to the close relationship between clinical information, morphological description and images. The availability of Whole Slide Images (WSI) enforces the development of scientific data management. Online managing and providing these heterogeneous data like natural language diagnoses, diagnostic findings, laboratory values or drugs information is a precondition for data driven hypotheses generation and, besides, easier collaboration among scientists and improved quality control.

Beside managing clinical and pathological data including images the availability of analyzing methods is a cornerstone for secondary usage of data and information in a research context. Scientists facing that big data need to connect data, knowledge and creativity in new manners to generate and accelerate innovation

In collaboration with the medical science partner OpEN.SC (Open Nephrology Science Centre) at Charité, Berlin [3] we develop a virtual research data environment. The primary application is developed as interface to OpEN.SC data. Later the tools will be applicable on other data too. To implement the tool we are designing a tool set based on ontology engineering and semantic data integration. Our information practices evaluation gives us orientation in the design process about the scientists' needs and habits: Our survey stated no explicit need for new software or problem solving support among physicians. The potential users create their ideas in exchange with colleagues or due to reading published papers. Our goal is to enhance this ideation approach and initiate it out of research data.

III. BUILDING A CREATIVITY FOSTERING KNOWLEDGE ENVIRONMENT FOR DATA EVALUATION

The methodology and strategy to create new ideas and concepts are changing: from hypothesis driven concept to explorative analysis. Since in data-driven science there is no initial search for specific data, data have to be presented in an interesting and encouraging way to dig deeper. Innovation, though, needs more than elaborated data representation and integration. If discovery of new knowledge shall be triggered from within the data, the data must be the starting point for scientific hypothesis generation.

In e-science the ideal type of an empirical research process–theory, hypothesis, data gathering, analysis, verification–should be expanded in the beginning. Hypothesis generation then is preceded by data exploration and the recognition of an Anomalous State of Knowledge (ASK) [4]. As stated by several approaches a scientific paradigm or the present knowledge of a researcher is the reference for remarking an anomaly which leads to problem awareness [5], [6: 252-254].

Since problem awareness and problem finding are the crucial steps in creativity which so far couldn't be replaced by computers (in difference to problem solving techniques), [7: 18] we are reasoning that a problem finding approach is most promising for generating innovation from research data.

The *DataCreativityTools for Innovation and Research (DCT)* will supply a process model to guide users in hypothesis generation and will supply also an information environment for enhancing users' knowledge for anomaly discovery. The tool is supposed to increase scientists' motivation in finding and analyzing information within the

data. Using the tool will improve knowledge discovery within the data that has been left unaccounted before.

IV. Designing a Process Model for Data Triggered Hypothesis Generation

The process model for data triggered hypothesis generation is compiled from cognitive models of creativity and information behavior. We believe that hypothesis generation should be modeled as a creative process itself, where the formulated hypothesis is considered as the creative product. Since "[p]roblem finding processes are initiated as attempts are made to organize and utilize information during the acquisition of further domain-specific content" [5: 213] they're entangled with information search.

Hence a model of information behavior should be regarded as subset of the creative process. We propose to conceptualize problem finding as iterative information work on an expanding knowledge base.

The functions implemented in *DataCreativityTools* will be built according to the stages of Cropley & Cropley's extended-phase model of creativity [6: 87-91]. For designing the information search loops as a feedback function within creative hypothesis generation we rely on Kuhlthau's Information Search Process (ISP) [8]. The system guides users through cyclical data exploring, information searches and hypotheses generation by encouraging task specific divergent or convergent thinking (Figure 1).

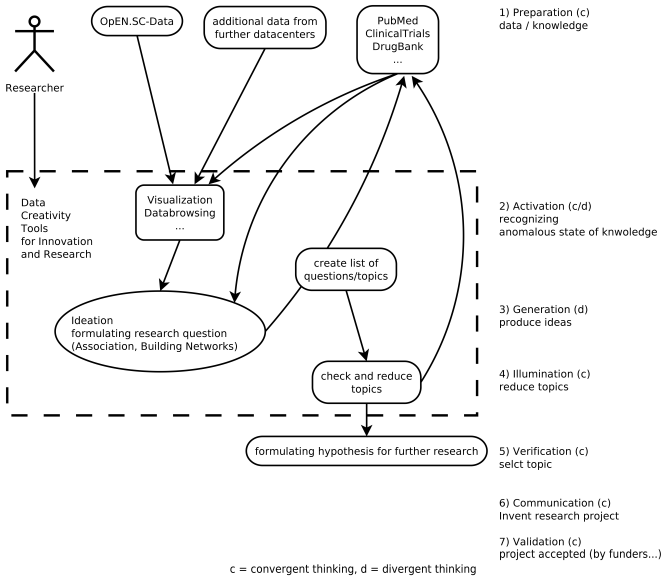

Figure 1. DCT functions and process mapped on Cropeley & Cropley's extended-phase model of creativity

Starting point for data exploration will be the OpEN.SC data. Further biomedical research data may be added to the system before or during the creative process. While browsing data attributes and visualizations (Activation) information from thoroughly selected sources will be integrated and displayed (Generation). As we know from Kuhlthau's ISP an information search is initiated when "a person becomes aware of a lack of knowledge or understanding" [8: 230]. In our case that means discovering an anomaly while exploring research data. Anomaly here refers to the knowledge landscape of the researcher. The interface will encourage formulation of hypotheses when an anomaly within the data is recognized. This could either be a personal lack of knowledge or discovery of a problem that hasn't been solved so far. The DCT will help to find out and to expand the users' knowledge bases. If the researcher has discovered a new problem, however, there simply will be no solution to be discovered. But how could one search for a specific gap of knowledge? The same problem (to integrate this discovered anomaly consistently into current knowledge or widely accepted theory) might be a matter of concern [9] elsewhere. Thus, the DCT will also integrate information search on publications and running projects. According to the ISP we expect and intend, that during the phase of "prefocus exploration" [8: 231] users' uncertainty about the problem will increase. DCT offer support in expanding the knowledge base and finding focus on specific nescience. Iterating searches will support ongoing shift between divergent and convergent thinking. Actually in our sense the ISP will be successfully finished when an anomaly within the research data is focused, all known facts are at hand and the search on publications and projects doesn't return anything but matters of concern.

Since searches are conceptualized and captured as research questions the user creates a list of research ideas while searching. Working later on that list for focus finding (Illumination) with the same DCT-configuration he or she is enabled to formulate innovative research hypotheses.

V. Building Knowledge Environment for Hypothesis Generation Support

Implementation of the model as a web based hypothesis generation support tool will rely on known techniques and sources.

DCT will consist from following components (Figure 2):

1) DCT Manager
2) Visualization
3) Data Browser
4) Domain Handling
5) Knowledge Manager
6) Hypotheses Manager

DCT Manager (1) is the central unit for component configuration and serves as interface node and process control unit.

Visualization (2) divides in two subunits for visualization of database semantics and of data values. SemaVis semantic visualization tool is supposed to be employed for this component. It allows data exploration in user customized manners [10: 29-32].

The Data Browser (3) allows exploring data through browsing properties simply to find out what data there are. It is based on an OpEN.SC specific developement and has to be adjusted for application on other domains.

Domain handling (4) requires a core data set which is the starting point for data triggered hypotheses generation. That is in our case the OpEN.SC data. There might be, however, more data centers out in the world that are of users' interest. Applying semantic web technologies, the Domain Extender is supposed to look for further research data. At best these data can be seamlessly integrated into OpEN.SC data. Since in most cases this is unlikely to work DCT will suggest exploring data in external environment.

The Knowledge Manager (5) serves as an interface to external sources. It supplies additional and task specific knowledge while data exploring. For the DCT prototype development the interfaces are classified into two types. Fact checking supports judging the data values and serves for gapping lacks of factual knowledge. As carried out in IV, finding matters of concern is the most important task in the process of hypotheses generation. Since it isn't sufficient to find an open question that is merely new to oneself, it has to be compared to the current work of others. Therefor the Research Field Explorer will commit searches within domain specific databases. De facto standards like PubMed will be selected for our application. Search results are evaluated statistically on criteria like number of publications per journal,

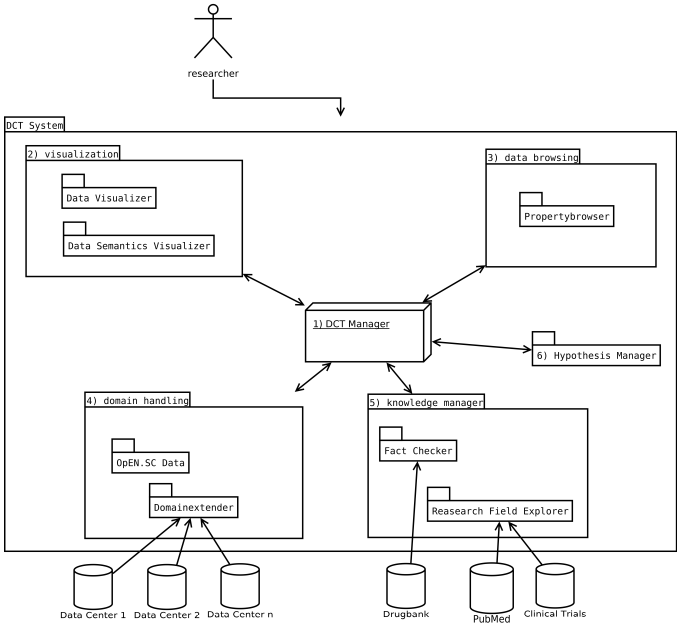

Figure 2. DCT system architecture

which is offered as additional information besides the list of hits. Further conceivable types for knowledge integration could refer to institutions or persons.

Every user activity on DCT is protocolled by the Hypotheses Manager (6). User input (e.g., on a certain property) is conceptualized as a potential research question and triggers a websearch carried out by components 4 and 5. The Hypotheses Manager also offers a search field for manual input of search terms. Thus, a list of potential research question evolves which implies certain hypotheses. Users are encouraged by the interface design to explicate these hypotheses. Since component 6 is integral part of DCT every verbalized hypothesis also will trigger a websearch for verification of the assumed new idea. The Hypotheses Manager will be a genuine *DCT* development.

VI. Conclusion and Outlook

The challenge we are facing when working on the vision of data-driven science is how to get beyond hypothesis driven science to a more flexible way shifting between data driven and hypothesis driven science. This task won't be based on a technical solution alone, but systems like *DCT* will support scientist who are willing to look at the data in new manners. They are envisioned as highly creative "brokers" between different domains concerning structural and domain specific knowledge. *DCT* will supply the construction material to the "brokers" when they're bridging the "structural holes [11]. This material consists from a generic model as presented in IV, sophisticated access to research data and semantic integration into additional data and information sources. Transferring *DCT* to further domains, though, depends on the availability of specific information sources and their stage of development of semantic information infrastructure.

References

[1] D.B. Kell., S.G. Oliver. "Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era." *Bioessays*, vol. 26, no. 1, pp. 99–105, 2004. doi: 10.1002/bies.10385

[2] A. Rector. *Barriers, approaches and research priorities for integrating biomedical ontologies. D 6.1.* Available: http://www.semantichealth.org/DELIVERABLES/SemanticHEALTH_D6_1.pdf, 2008 [Jul. 19, 2012].

[3] T. Schrader, B. Rudolph, M. Dietel, M. Beil, T. Schaaf, D. Schmidt, G. Lindemann. „The Open European Nephrology Science Centre (OpEN.SC) - an information service center for kidney diseases and transplantation." *Pathology Research and Practice*, vol. 203, no. 5, p. 314, 2007.

[4] N. Belkin. "Anomalous State of Knowledge." in *ASIST monograph series*, *Theories of Information Behavior*, K. Fisher, S. Erdelez and L. McKechnie, Eds. 2nd ed, Medford, NJ: Information Today, 2006. pp. 44-48.

[5] S.M. Hoover, J.F. Feldhusen. "Scientific Problem Solving and Problem Finding: A Theoretical Model." in *Creativity research, Problem finding,*

problem solving, and creativity, M.A. Runco, Ed., Norwood N.J: Ablex Pub. Corp, 1994, pp. 201–219.

[6] A. Cropley, D. Cropley. *Fostering creativity. A diagnostic approach for higher education and organizations. Perspectives on creativity research.* Cresskill, NJ: Hampton Press, 2009.

[7] M.A. Runco. *Creativity. Theories and themes ; research, development, and practice.* Amsterdam: Elsevier, 2007.

[8] C.C. Kuhlthau. "Kuhlthau´s Information Search Process" in *ASIST monograph series*, *Theories of Information Behavior*, K. Fisher, S. Erdelez and L. McKechnie, Eds., 2nd ed, Medford, NJ: Information Today, 2006. pp. 230–234.

[9] B. Latour. "Why Has Critique Run out of Steam? From Matters of Fact to Matters of Concern." *Critical Inquiry*, vol. 30, no. 2, 2004. pp. 225–248. doi: 10.1086/421123

[10] R. Schäfer, T. Becker, C. Burghart, K. Nazemi, P. Ndjiki, T. Riegel. „Basistechnologien für das Internet der Dienste." in *Internet der Dienste*, L. Heuser, W. Wahlster, Eds., Berlin, Heidelber: Springer, 2011. pp. 19-40.

[11] R.S. Burt. "Structural Holes and Good Ideas." *American Journal of Sociology*, vol. 110, no. 2, 2004. pp. 349–399. doi: 10.1086/421787