Scientific Discovery: A View from the Trenches

Catherine Blake[1], Meredith Rendall[1]

[1] University of North Carolina at Chapel Hill, School of Information and Library Science,
214A Manning Hall, Chapel Hill, NC, USA 27599
{cablake, mbr} @email.unc.edu

Abstract. One of the primary goals in discovery science is to understand the human scientific reasoning processes. Despite sporadic success of automated discovery systems, few studies have systematically explored the socio-technical environments in which a discovery tool will ultimately be embedded. Modeling day-to-day activities of experienced scientists as they develop and verify hypotheses provides both a glimpse into the human cognitive processes surrounding discovery and a deeper understanding of the characteristics that are required for a discovery system to be successful. In this paper, we describe a study of experienced faculty in chemistry and chemical engineering as they engage in what Kuhn would call "normal" science, focusing in particular on how these scientists characterize discovery, how they arrive at their research question, and the processes they use to transform an initial idea into a subsequent publication. We discuss gaps between current definitions used in discovery science, and examples of system design improvements that would better support the information environment and activities in normal science.

Keywords: Socio-technical, information behaviors, knowledge discovery.

1. Introduction

As scientists, we often find the magic that surrounds a scientific discovery captivating. How could we forget Kekulé's dream of a snake eating its tail that revealed to him the elusive benzene ring structure; or the contamination in Fleming's lab that lead to the profoundly important discovery of penicillin? Kuhn would refer to these landmark discoveries as "scientific revolutions"[7]. Although revolutionary discoveries are powerful ways to attract new-comers to a field, or to keep ourselves motivated, scientists spend much of their time on day-to-day activities or what Kuhn would call "normal" science.

The underlying premise of our research is that capturing the day-to-day activities used

by scientists as they develop and verify hypotheses will provide both a glimpse into the human cognitive processes surrounding discovery and into the complex socio-technical environments in which successful discovery tools will eventually be embedded. Simon, the unequivocal father of discovery science, and his colleagues, stated that "Discovery systems which solve tasks cooperatively with a domain expert are likely to have an important role, because in any nontrivial domain, it will be virtually impossible to provide the system with a complete theory which is anyway constantly evolving" [11]. The need for additional user involvement in discovery systems is echoed by Langley who predicted that "as developers realize the need to provide explicit support for human intervention, we will see even more productive systems and even more impressive discoveries" [8].

Our position is that in addition to the difficulty in encoding prior knowledge and the need to design more interactive computational discovery systems, we need to understand the broader context in which science takes place to ensure the habitual adoption of discovery systems.

If adoption is one of the criteria of a successful discovery system, then it is critical that that we understand the work environment in which the discovery system will eventually be embedded. In this paper, we explore the processes used by experienced faculty in chemistry and chemical engineering. We focus in particular on how scientists define discovery, how they arrive at their research question and the processes used to transform research questions into published manuscripts. To frame our conversations, we interviewed 21 experienced scientists, using the critical incident technique[3] that employed two papers where each scientist was principal investigator and one paper they considered seminal to the field. Although we concur with Simon, et al., that a single publication captures only a highly circumscribed problem [11]; a published manuscript seems a natural level of analysis because we refer to work at this level through citations and because institutions measure the productivity of a scientist in terms of the number and quality of publications for promotion. Langley for example used publication as the "main criterion for success" of computer-aided scientific discoveries [8].

2. Research design

Kuhn (1996) described discovery as "an inherently complex task". Although much of the previous research in discovery science has emphasized automated discovery, we seek to develop a model of the reasoning processes that surrounds a scientist's day-to-day activities. In this study, we focus on the processes used to develop a research question (hypothesis development) and to transition from that question to a final published manuscript (hypothesis verification), activities that consume much of a scientist's time. Our strategy is to conduct and record a series of interviews with experts in chemistry and chemical engineering. Chemistry has been a fruitful area for automated discovery systems, such as in MECHEM [14] and FAHRENHEIT [16].

2.1 Recruitment

The interviews reported in this paper were the first step in a two-year Text Mining Chemistry Literature project funded by the NSF in conjunction with the Center for Environmentally Responsible Solvents and Processes (CERSP). The primary investigators for the Center were strong supporters of the proposed interviews and sent the initial invitation for participation. Of the 25 scientists in the Center 21 members participated, resulting in a response rate of 84%. Inclusion in the interview process was determined solely by whether the participant was currently or had recently been involved in chemistry research. To ensure that subjects were familiar with the scientific discovery process, we required that they had previously published at least three articles and written one successful grant. Semi-structured interviews were conducted, recorded, transcribed verbatim, and analyzed during the Spring and Summer 2006.

2.2 Methodology

After approval by the Institutional Review Board of the University of North Carolina at Chapel Hill, we conducted the interviews in each scientist's office. We confirmed their title, role, and affiliations and then asked targeted questions regarding their definition of a discovery, the factors that limited the adoption and deployment of a new discovery, the processes they used to come up with a research question and verify the hypothesis. We also asked questions about the role scientific literature plays during their discovery processes and about their information use behaviors with respect to information overload, but the latter two themes are beyond the scope of this paper.

Data was collected using the critical incident technique [3] based on three articles or research grants: two the scientist wrote (one recently completed project and one project of which they were particularly proud), and a third they considered seminal to the field. For those scientists who were unable to provide these articles, we recommended frequently cited articles that we identified from a corpus of 103000 full text chemistry articles.

In addition to our interview questions, we asked scientists to reflect on the process they used for each of the two papers they had written. We provided the following activity cards: reading, thinking, online searching, books, journals, experimenting, analyzing, writing, discussing, and organizing; we asked the scientists to organize the cards in a way that reflected the scientific process used in the first paper (we also provided repeat cards and invited scientists to add steps). We then asked them to review the cards to ensure that the process reflected their second and subsequent research processes. We asked they simultaneously describe the process. We recorded and subsequently transcribed the entire interview and used grounded theory[5] to identify reoccurring themes.

3. Results

The average duration of the recorded interviews was 52 minutes. Of the 21 participants,

half provided articles, and we suggested articles for the remaining 11 participants. None of the scientists provided a research grant. Figure 1 contains a summary of the participants including their ID letter that we will use in the remainder of this paper. Interviews were transcribed (including those that were not recorded) and analyzed using NVivo 7 [9]. Once collected the data was analyzed using a grounded theory approach [5].

ID	Interview (mins)	Title	Area of Research	Experience (yrs)
A	67	Director	Biochemical Engineering	32
B	55	Assistant Professor	Colloid Science and Engineering	10
C	51	Associate Professor	Polymer Design and Synthesis	12
D	51	Professor	Semiconductor Surface Chemistry	34
E	43	Professor	Polymer Chemistry	16
F	60	Professor	Nanoelectronics and Photonics	10
G	50	Professor	Electronic Materials Synthesis	26
H	59	Director	Polymer Design and Synthesis	35
I	39	Assistant Professor	Colloidal & Macromolecular Physics	14
J	61	Professor	Nanoelectronics and Photonics	36
K	58	Associate Professor	Bioorganic Chemistry	7
L	44	Professor	Rheology	13
M	41	Professor	Organometallic Chemistry	31
N	54	Professor	Polymer Theory	23
O	41	Professor	Electrochemistry	46
P	56	Professor	Synthetic Organometallic Chemistry	37
Q	5*	Associate Professor	Surface and Interface Polymers	10
R	†	Associate Professor	Polymer Thin Films	20
S	53	Director	Polymer Synthesis	16
X	56	Professor	Neutron Scattering	35
Y	33	Professor	Chemical Reaction Engineering	40

Fig. 1. Summary of participants. *This time reflects a partial recording. † Declined permission to record the interview.

3.1 Definition of a scientific discovery

The evaluation of a discovery system is problematic without a definition of scientific discovery. One of the most cited discovery definitions is borrowed from Knowledge Discovery in Databases (KDD), which has been defined as “the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data”

[2]. Valdés-Pérez revised this statement to, “Discovery in science is the generation of novel, interesting, plausible and intelligible knowledge about the objects of study” [13]. In addition to asking directly, “What is your definition of discovery?”, the seminal papers provided a framework for the discovery conversations. Nineteen of the 21 participants indicated the discovery characteristics, which fell into five key themes: novelty, building on existing ideas, a practical application, experimentation and theory, and simplicity.

3.1.1 Novelty

It was of little surprise that novelty was the most common theme surrounding discovery (11 out of 19 cases). Terminology comprised “not previously seen” (M); “new insight” (G and M); “obviously novel and new and doesn’t exist in the literature” (L); “finding something new and unexpected” (P); “learning something that hasn’t really been well understood before (G)”; and “it [discovery] opens the door to exploration” (O).

It was interesting that in addition to new materials and substances, scientists provided examples of new transformations and processes. For example, scientist J said, “Novelty meaning novelty in what you are going to look at, how you are going to look at it, what you expect.” Scientist R best characterized this difference between the actual output and the way that the problem had been characterized.

> You need to understand there are two different types of seminal papers. There are the some people working in certain fields, accumulating a lot of data with current models. They run experiments which are 80-90% of the time in agreement with the current models. 10-20% of the time they disagree with the model. At a certain point, these people reach a critical point when they can support a claim that the model is wrong.
>
> At this point, they put forth a new point of view/model, which raises the level of science. Perhaps this provides new language, ways of explaining what is going on, and perhaps develops a new/tangential line of science.
>
> Then there are other seminal papers that discuss a fundamental topic in a new light. For instances, hundreds of groups are working on the same topic, coming from the same angle, all trying to make it through the doorway first. But there are different approaches, and if you can find a different approach, you can find another door that unlocks the mystery everyone else is trying to solve simply because you took a different tack than any of the other groups working on the issue.

Combined with the expectation of novelty and the need for something unexpected, scientist P said, “A discovery is more important than something partially anticipated”(Q), whereas another scientist thought discoveries could be “planned or serendipitous”.

3.1.2 Building on existing ideas

Ironically, the second most frequently occurring theme surrounding discovery was an improved understanding of an existing mechanism (7 out of 19 cases). Several scientists doubted that anything completely new existed. “Even supposedly the most creative people…I don’t think things are cut from a whole cloth anymore. I think there aren’t any more cloths without big holes anymore” (I); “Everything has precedent, in my opinion” (M); “from my standards, one has only less than ten completely new ideas in their lifetime, and so, most of the time you are sort of doing some modifications on a new idea or something” (N).

3.1.3 Practical application

On par with the number of scientists who suggested building on previous ideas was a discussion for a practical application of the discovery (7 out of 19 cases), which is akin to the “useful” requirement proposed by Fayyad and Valdés-Pérez. However, in contrast to explaining what was known our data revealed a tension between differentiating discovery and practice and the need for a practical or commercial application. “I’d much rather work in the discovery mode: we discover something, get enough data to publish it, and think about how might this be commercialized, what would we need to make it commercializable, who might be interested in commercializing it. But that’s kind of a secondary question. The primary question is ‘what did we learn?’ “What new insight can we bring to the field?” Rather than is this going to make my funding agent happy” (G).

Scientist M’s experience when working for a funding agency captured the tension between practice and theory “Gee, I got tired of reading that this was the first this and the first that.” And I was, dependent on my outlook how frank I am, but I said, “This is the first time in April that I’ve reviewed a paper for [removed to maintain anonymity]!” But who cares? So there are lots of firsts that are wearing the shirt of a different color or making a new isorun that nobody cares about. I don’t put those in a discovery category. On the other hand, NSF likes to try to support research that is going to lead to discoveries that will be transformational and start new areas of research. Those are hard things to do and don’t happen very often.”

3.1.4 Experimentation and theory

Scientists grappled with the duality of being both a theorist and an experimentalist. “… as an experimentalist, I treat analyzing data as ‘let me try to decide whether or not what I have measured is real before I get too excited about it’. It goes on and on, analyzing in the context of discussing and thinking about what’s occurring. I think when you try to do experiments, you have to disconnect yourself instead of getting all excited about results and thinking you have found something when you might not have” (I). This tension

between developing theories that truly reflects the available evidence is a hallmark of a good scientist.

In one case, the scientist had not yet published his theory because they were "waiting for more proofs" (N). In another, the scientist reflected on the importance of experimentation to test their theoretical knowledge, "There's this added level of, almost, engineering that we build a system to see if we can mimic what nature does. If we can mimic what nature does then that tells us that we do understand it as well as we think that we do" (K).

Valdés-Pérez observed the theoretical motivations in chemistry when he stated that chemistry is, "more of theory-driven than many of the recent tasks addressed by data-driven discovery research" [15]. Our results suggest that scientists require discoveries to include both a theoretical foundation, and supporting experimental evidence.

3.1.5 Simplicity

Fayyad, et al,'s discovery definition required that the process be non-trivial. Stemming from the conversation of the importance of "a good description" was the need for simplicity which is typified by the following comments: "A discovery doesn't have to be something that is very hard. It could be something very simple that you can get to work. It doesn't have to be tedious work or years spent. Some people to see something simple might say it is way too simple, but as long as it is an elegant thing and hasn't been thought out already; I think that's fine. It's a discovery" (L), and "I was trying to make as simple as possible understanding of how this long molecules move sort of like spaghetti. … So this was my attempt to make to simplify to the minimum possible description simple as possible description of this very complicated problem" (N).

3.2 Arriving at a research question

Discovery systems have been used to confirm or refute an existing hypotheses or to generate hypotheses. Our second goal is to explore how scientists arrive at their research questions. We conjecture that an improved understanding of this process will lead to improvements in automated discovery systems that assist in hypothesis generation.

One scientist noted that crystallizing the question and refining the problem were the fundamental elements of success in his projects, saying, "From my point of view, to get a good research question you have to define the problem properly. If you don't define the problem, you cannot do anything. Once you define the problem clear enough, you can, if the person is smart, you can find a solution. The difficulty of not being able to find a solution is just because you haven't crystallized and clearly posed the problem" (N).

All 21 scientists described how they arrived at their research question. Their sources fell into four key themes: discussion, previous projects, combining expertise and the literature.

3.2.1 Discussion

Discussions with colleagues and students were the most heavily cited source of idea generation (14 out of 21 cases). Scientists consulted with colleagues one-on-one and during Center presentations to clarify ideas in a paper or to investigate new avenues of research. Scientist I best captured the importance of discussions: “It’s just like anything else, each time you start explaining something to somebody you realize how much you know and how well you understand it. To me that’s discussing things, and actively discussing the data in general.”

Several scientists acknowledged that the level of cooperation depended on the degree to which they were in competition or collaboration with another research group. The greatest cooperation occurred between different areas of science rather than within one genre, unless the scientists were collaborating, in which case there was a regular stream of information.

One scientist identified the importance of informal friendships, where he spent “40 days together over a two-year time period doing some DARPA work and a lot of bus rides, finding out what each other does. He does that field. I simply asked the question. What are the [removed to ensure anonymity] problems he has? He was able to spell them out and I said, we can fix that. He said no, people have been trying to fix that. I said, give us a shot, and we did. That’s how it started” (S). Another scientist emphasized the importance of external conversations “Talking to people outside my field is something I would always do. I would never consider it a waste of my time to talk with people” (Q).

Conferences provided another opportunity for discussions. Attending presentations and participating in conversations were integral to the generation of new ideas and collaborations. One scientist noted that ‘hearing about other stuff’ (K) provided the source of their ideas, and incorporating ideas and processes into their research from related areas was beneficial. Discussions also arose from lectures, seminars, and interactions with visiting scholars.

3.2.2 Previous projects

It is of little surprise that on-going projects also determined new research questions (13 out of 21 cases). Though our scientists acknowledged the evolution of ideas, many remained in the same general area as their graduate or post-doctoral work, stating: “I’ve been doing that since 1972” (D) and “This hard earned body of knowledge is what must determine future projects, whether mechanistic or experimental” (L).

Several of the discoveries identified by these scientists stemmed from experimental “mistakes”, which one scientist challenged with, “Is it really wrong? No. It might be interesting” (E). Another stated that “There are times when an investigation finds itself off course and heading in an unanticipated direction, which may be for a variety of reasons including the original idea looking less and less promising to the unexpected outcome is very exciting and potentially a new area of science” (P).

One scientist called his group the ‘follow your nose’ group, meaning they based their next set of experiments on the data from that day’s experiments, asking what had they learned, what were the obstacles, and what did that mean?” (E). Deciding to follow unexpected results or not depended on a matrix of situational elements, but the key elements seemed to be the purpose of the research and the perceived likelihood of success. If the purpose of the project was a mechanistic understanding of a system, then research groups were less likely to follow their noses. In some cases, the outcome of a project was the need to develop a new tool or technique. At least two of the scientists stated they had the only tool or technique in the world.

3.2.3 Combining expertise

A re-occurring theme in arriving at a research question was collaborations with colleagues from different areas of chemistry. Scientist N reflected the international dimension of those collaborations on his visit to a lab in France “They had some very strange results they couldn’t explain, and on the other side we were working on this different part of the same problem. So the two things clicked. That’s how, from a discussion in France and our research at EKC, this idea came that this must be new.” This theme relates closely to discussions in general (see section 3.2.1); however, the differing backgrounds and geographical locations might have implications to a discovery system that supports these behaviors.

3.2.4 Reading literature

Chemists read literature more articles per year and spend more time reading compared with scientists in other fields[13]. It is no surprise then, that literature was included as a source for new research questions. The scientists cited discrepancies and errors in articles as a strong motivator, with one scientist stating that he had to conduct the research because he found the article “chemically offensive” (E).

3.3 The “normal” scientific process

Our third goal was to model the process used by scientists to transform their initial research hypothesis into a published manuscript (hypothesis verification). We have created transition and co-occurrence models that reflect the process diagrams collected in this study; however due to space limitations in this paper, we provide recommendations that would close the gap between existing computational discovery system designs and the functionality required to support ‘normal’ science.

Figure 2 shows one of the process diagrams (selected at random). The repeat cycles for cards on left and on the bottom right of the image were typical of process diagrams

collected in this study. The most frequent iteration occurred between experimentation and analysis. A discovery system that integrated activities would better support iteration than existing stand-alone designs. For example, a discovery system that integrated the data collected during experimentation with analysis tools could then identify patterns that agreed with or refuted the scientist's current experimental evidence.

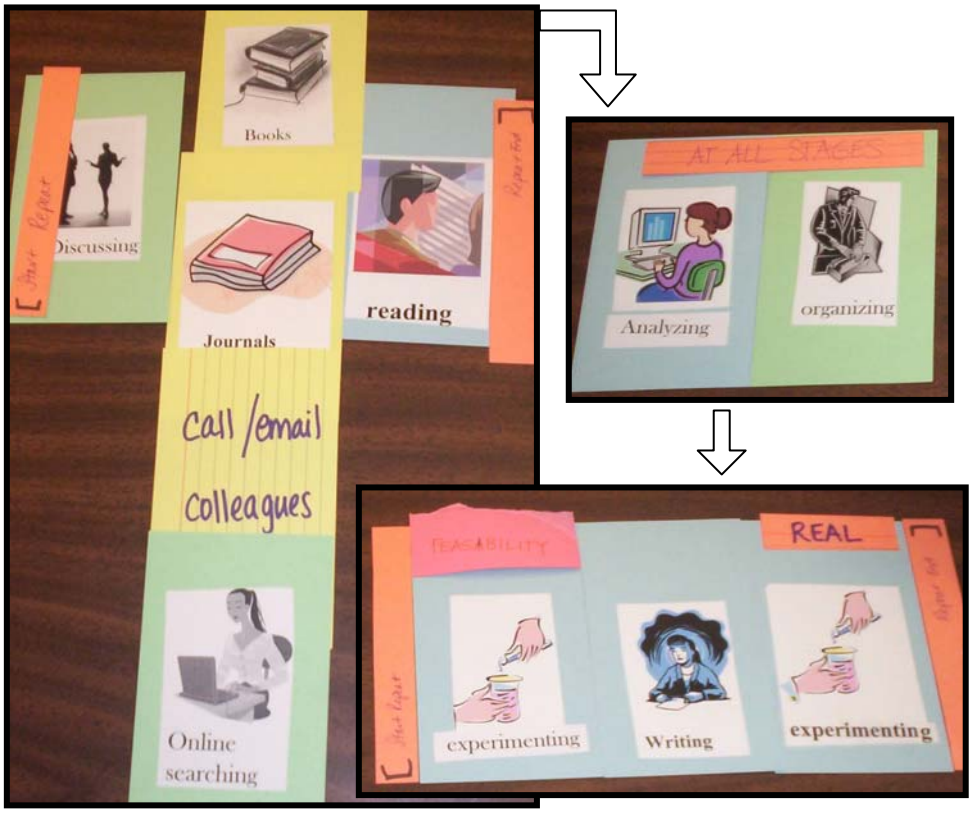

Fig. 2. An example of the process used to write a published manuscript.

Consider again Figure 2, where the scientist's first iteration includes books, journals, online searching and reading. Perhaps the biggest gap between our observations of a scientist's day-to-day activities and existing computational discovery systems is the need to incorporate literature into the system. While discovery systems typically distinguish between structured and unstructured (aka data mining versus text mining), our observations suggest that data in both representations are critical to the process of verifying or refuting the validity of a new hypothesis. A discovery system that incorporated both could provide scientists with related articles and summaries of how

previous work compares with current experimental findings. Such a system could also ease the transition from reading and analysis to writing by providing support for citation management.

Our data collected during this study suggests that scientist do not arrive at the discovery process with no a priori expectations. Instead, they start with a hypothesis projection, "the purely conjectural proliferation of a whole gamut of alternative explanatory hypotheses that are relatively plausible, a proliferation based on guesswork - though not 'mere' guesswork, but guesswork guided by a scientifically trained intuition."[8] To support this behavior, a discovery system should identify patterns that relate to a user's previous studies. The characterizations of discovery (see section 3.1.2) also reflect the need to relate current findings to previous work.

The process diagrams suggest that the scientific process is inherently a social endeavor, yet few discovery systems support interactions between scientists, or between scientists and students. Although mechanisms such as email (as shown in Figure 2) and instant messenger are available to scientists, discovery system designers have yet to incorporate these technologies into their design. In addition to real time communications, a discovery system could support collaboration by enabling scientists to share data, annotations, and manuscripts.

These findings are consistent with a previous study of scientists in medicine and public health [1] and with existing cognitive science perspectives, such as personal construct theory, which emphasizes the importance of inconsistencies among information artifacts and between an information artifact and a user's mental model [6]. Kelly suggests that inconsistencies force a user to discard information that threatens their existing mental models or formulate a tentative hypothesis. From the personal construct theory perspective, our study deemphasizes the early stages of confusion, doubt, and threat focusing on hypothesis testing, assessing, and reconstruing. Gardner's cognitive model suggests that synthesis plays a pivotal role in information interactions [4]. He states, "the organism ... manipulates and otherwise reorders the information it freshly encounters – perhaps distorting the information as it is being assimilated, perhaps recoding it into more familiar or convenient form once it has been initially apprehended." We observed re-ordering as several scientists used the organizing activity card in their process diagram.

5. Conclusions

This study explores activities conducted during what Kuhn would call "normal" science, including how scientists arrive at a research question (hypothesis development) and the processes used to transform a research question into a published manuscript (hypothesis verification). Such studies are critical if we are to design discovery systems that "solve tasks cooperatively with a domain expert" [10].

The most frequent themes surrounding the definition of discovery were novelty, building on existing ideas, practical application, conflict between experimentation and theory, and the need for simplicity. Such themes provide discovery systems designers with

new criterion by which to measure the success of their automated discovery systems.

This study reveals the iterative nature of the scientific process, and the inherently social context in which normal science place. It is only by embedding computational methods of discovery into such an environment that we can ensure their habitual adoption.

Acknowledgments. Nancy Baker contributed to earlier discussions surrounding the activity cards and methodology used to collect the process diagrams from scientists. This material is based upon work supported in part by the STC Program of the National Science Foundation under Agreement No. CHE-9876674. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] Blake, C. & Pratt, W. (In Press, 2006). Collaborative information synthesis I: A model of information behaviors of scientists in medicine and public health. To appear in *Journal of the American Society of Information Science and Technology*.

[2] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. *Advances in Knowledge Discovery and Data Mining*. AAAI Press, 1996.

[3] Flanagan, J.C. The Critical Incident Technique. *Psychological Bulletin*, *51* (4). 327-359, 1954.

[4] Gardner, H. *The mind's new science : a history of the cognitive revolution.* Basic Books, New York, 1985.

[5] Glaser, B.G. and Strauss, A.L. *The discovery of grounded theory. Strategies for qualitative research*. Chicago, Aldine Pub. Co, Chicago, 1967.

[6] Kelly, G.A. *A theory of personality: The psychology of personal constructs*. Norton, New York, 1963.

[7] Kuhn, T.S. *The Structure of Scientific Revolutions*. The University of Chicago Press, Chicago, 1996.

[8] Langley, P. The Computer-Aided Discovery of Scientific Knowledge. *Proceedings of the First International Conference on Discovery Science*, *1* (4). 423-452, 1998.

[9] Rescher, N. (1978). Peirce's philosophy of science critical studies in his theory of induction and scientific method. Notre Dame, London: University of Notre Dame Press.

[10] QSR International Pty Ltd. www.qsrinternational.com/products/productoverview

[11] Simon, H.A., Valdés-Pérez, R.E. and Sleeman, D.H. Scientific discovery and simplicity of method. *Artificial Intelligence*, *91* (2). 177-181, 1997.

[12] Tenopir, C., King, D. W., Boyce, P., Grayson, M., Zhang, Y., & Ebuen, M. (2003). Patterns of journal use by scientists through three evolutionary phases. D-Lib Magazine, 9, 1-15.

[13] Valdés-Pérez, R. Principles of human computer collaboration for knowledge discovery in science. *Artificial Intelligence*, 107 (2). 335-346, 1999.

[14] Valdés-Pérez, R.E. Some Recent Human-Computer Discoveries in Science and What Accounts or Them. *AI Magazine*, 16 (3). 37-44, 1995.

[15] Zytkow, J.M., Combining many searches in the FAHRENHEIT discovery system. in *Proceedings of the 4th International workshop on machine learning*, (San Mateo, 1987),281-7.

[16] Zytkow, J.M. Integration of knowledge and method in real-world discovery. *ACM SIGART Bulletin*, *2* (4). 179-184, 1991.