Innovators Marketplace on Data Jackets for Externalizing the Value of Data via Stakeholders' Requirement Communication

Yukio Ohsawa, Chang Liu, Yoshitaka Suda, and Hiroyuki Kido
School of Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku Tokyo 113-8656 Japan
ohsawa@sys.t.u-tokyo.ac.jp

Abstract

We here show an approach of the market of data, toward the progress of three core challenges of this workshop. In this market, stakeholders of essential problems in businesses, sciences, politics, and healthcare, can communicate to externalize and share the value of datasets, motivated by the desire to collect/provide data in a reasonable condition, e.g., for a reasonable price. Here we show a basic research about one way to realize such a market, that we call the Innovators' Marketplace on Data Jackets, where the content of each dataset may be hidden due to constraints that are strict in healthcare, but the digest of the dataset should and can be disclosed for expressing the latent value of the data. Based on those digests called Data Jackets (DJs), relations among datasets are visualized for aiding stakeholders' creative communications for expressing latent requirements and solutions for satisfying them. Here we show how requirements of users play a role in creating use scenarios of data and proposing missed DJs.

Introduction: The Market of Data

The market of data is not about data mining for marketing, but means a social environment where data are reasonably dealt with, i.e., sold, opened free, or shared after negotiation. We already have the market of data on the earth, such as the on-line ones (e.g., Windows Azure Marketpalce and KDnugget in the reference), where a large number of data digests are exhibited. Their outlook may be similar to data catalogue as in CKAN (see reference), but they do not always open the data free – prices are assigned to each dataset in advance or via negotiation between the provider and user. However, there is no guarantee that the user or the provider can choose the most useful from a number of datasets, or from a number of possible conditions to sell/buy/use a dataset. Such a choice can be realized by comparing one dataset with others, or discussing scenarios to combine, analyze, and use data. Therefore we should collect various datasets and people in order to realize creative communications where expertise and experiences can be exchanged and combined without confusions. From such a communication in heterogeneity, we expect to combine "quantitative data and qualitative information for improving" healthcare, as called in the workshop CFP (Kido and Takadama 2014).

An important point for healthcare system is that our goal is to have each doctor, patient, or owner (i.e., potential provider) of data, who are potentially relevant to healthcare, share one's own or others' data without fearing the loss of benefits. Generally, we define stakeholders as people who can be regarded to be involved in the process to solve a problem. We aim to create a social system where stakeholders can share data for externalizing and solving problems they potentially share. In the society, datasets should be priced reasonably on the negotiation among data owners including patients, doctors who find strategies for treatment from data, and brokers who suggest how the datasets may be beneficial to each stakeholder. In other cases, the data provider may decide to expose the data free i.e. as open-source, if the communication concludes that the analysis may lead to a world-wide progress of health care systems. Thus, the market of data means a place where the value of data and scenarios for combining and using data are communicated, externalized, and shared.

Furthermore, data analysts – who may be called data scientists – often desire to import techniques from each other. Such techniques are not easy to learn from others dealing with different kinds of data, since the similarity of datasets is not always obvious. For analogical reasoning in the analysis of the target data, it is an essential step to acquire structural model of the latent causality in the target domain and match it with the latent structure underlying previously analyzed data (Gentner 1983). Computational methods for learning such useful structures extending such a recent work in (Bollegala et al 2013), are keenly desired as well as co-creative communications among experts e.g. medical doctors and inspectors in the case of blood test data (Ohsawa et al, 2006).

Innovators Marketplace on Data Jackets

Here let us review the abstract of our approach toward realizing such a market of data as we desire, with revising the original proposal of Innovators Marketplace on Data Jackets (IMDJ: Ohsawa et al 2013). The basic idea of IMDJ comes from what we and salesclerks do in shopping stores of media with data such as movie DVDs and music CDs, where we find a poster showing "the shelves exhibit only jackets, not the content, until you pay for the disk." That is, only quite superficial information i.e., digests that are free from any risk to lose commercial value, tend to be shown open to the public. On the other hand, the content should be hidden in order to reduce the risk – otherwise the details may be copied and used free by anyone who does not pay, or analyzed by rivals for their own benefits.

Such a conservative policy of data closure may look suppressing to innovations. However, the idea here is really useful when we desire a social environment where data are accessible *only* from those who seek, and where each data can be valued or priced reasonably with a logical understanding of possible use scenarios of data. This is the essential idea in the basis of IMDJ, in which each data owner takes part in the market of data with filling in and disclosing a *data jacket* (DJ in short), that is just a small piece (card-size) of information describing the digest of one's own data. Based on the collected digests, negotiations of data owners and stakeholders who may become interested in the data are initiated. In contrast to big data, which are hard for small firms to collect and hard to manage even for huge companies, DJs are easy to collect and used for discovering possibilities to link datasets to each other. For example, a DJ provided by a single participant about his/her family member or friends of whom privacies should be closed, can be disclosed so that someone taking part in IMDJ may find a scenario to combine it with other DJs for discovering new patterns of social activities. Thus, the data behind each DJ can be evaluated, on the small pieces of information.

Innovators Marketplace as the basis for IMDJ

IMDJ essentially follows the Innovators Marketplace. The Innovators Marketplace® (trademark registered by author: IM hereafter, see Ohsawa and Nishihara 2012 for details) is a process for innovative collaboration consisting of pre-process, the game called the Innovators' Market Game (IMG), and the post process. In the pre-process, existing pieces of knowledge are collected from and by stakeholders of a topic or a problem given in an abstract form such as "how can we protect the city from illness?" These pieces are given to IMG, where innovative couplings of multidisciplinary stakeholders' knowledge are proposed, in communicating similarly to in the real market. Finally, in the post-process, the ideas created in IMG are further developed into feasible plans of business, via logically grounding to real conditions in business.

The remarkable novelty of the IM process is in the main part IMG, where a game board is made visualizing possibilities to combine element pieces of knowledge provided by stakeholders taking part in the game. The game board is made mostly using KeyGraph®, our original tool used in real business sites for visualizing correlations among items (words, products, services, technologies, etc).

In KeyGraph, novel items or rarely used knowledge are visualized with highlight if positioned on bridges between clusters of items that appear and co-occur, i.e., appear close to each other, frequently in the data. About details of KeyGraph see (Ohsawa, Benson, and Yachida 1998, Hong et al 2006, Fruchter, Ohsawa, and Matsumura 2005, and Llor et al 2006, etc). Its extension can be seen in (Ohsawa 2005). In comparison with the prediction of rare events, studied quite long since distinguished studies as in (Joshi 2002, Weiss and Hirsh 1998), KeyGraph aids the externalization of users' subjective interest in items or events that have not been highlighted so far, but are useful for catching up with the emergence of trends and for creating action scenarios. In other words, KeyGraph and its extensions work for urging *chance discovery*, that means the discovery of events significant for humans' and agents' decision making (Ohsawa and McBurney 2003).

In (Ohsawa and Nishihara 2012), we linked chance discovery to users' sense making (Dervin 1983), i.e., to give meaning to each event in daily-life experiences by the interpretation of data. For urging this effect, the interaction among elements on the game board and in the knowledge of participants should be conducted with looking at the board, because the meaning of events for solving a social problem may not be able to be given by the sheer thought of a human nor by knowledge provided before communication starts. For this purpose, the elements are printed on small cards, so that participants can pick elements in hands and exchange with others via trading with communicating in IMG. By this interaction, combinatorial creativity i.e. the participants' ability to create a new idea from combination of existing knowledge, is reinforced. The game starts, therefore, with a prepared set of cards, on which titles and summaries of existing knowledge/technologies are printed.

The players of IMG are classified into *inventors* and *consumers* (Figure 1), although sometimes one participant plays the roles of both. Inventor's most important task is to combine the prepared cards to create and present ideas about new products/services. In combining, they refer to lines on the printed game board, to create ideas. Cards linked via lines tend to be easy to combine, whereas other combinations are hard but may trigger the generation of ideas of high novelty. Inventors may propose or accept

collaborations with negotiating the condition for sharing or cards or ideas. Consumers buy an idea presented by some inventor only if buying is expected to improve the quality of their quality of life (QOT), on negotiating with the inventor who created the idea in order to set the price. In this negotiation, *consumer*s may evaluate or criticize the ideas, or present requirements that inventors are desired to satisfy by creating ideas. Each inventor's idea and each consumer's requirement should be spoken orally, and also written on a sticker to be put close to relevant elements on the game board. As a result of IMG, the inventor rewarded with the largest amount of money becomes the winning inventor. Because toy money is used in IMG rather than real USD or JPY, we regard the competition here as a preparation for the real trading where real money and data are exchanged. And, the consumer evaluated the highest by other participants becomes the winning consumer, according to the presentation about purchased ideas and the consequent improvement of his/her own life quality.

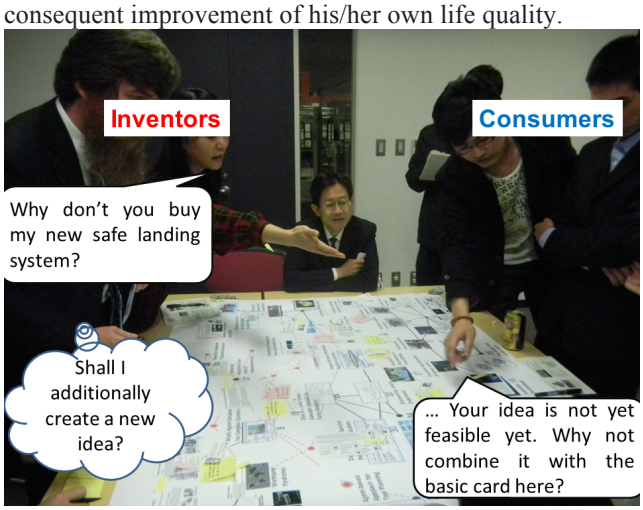

Figure 1 A gaming scene of IMG, in the process of Innovators Marketplace®.

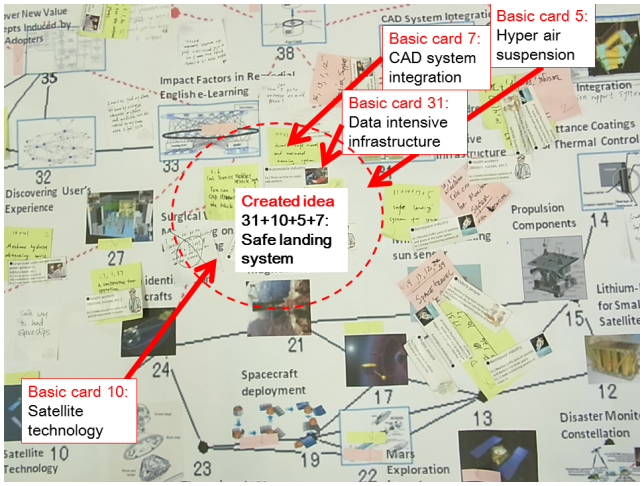

Figure 2 A result of IMG

Figure 2 shows a result of IMG in year 2010, where the problem was set as how to inventing service systems by using technologies, in Workshop on Innovating Service Systems (ISS2010), organized by The Japanese Society of AI. Here, 40 element technologies, prepared in advance, were mostly on architecture, space technologies, and computer technologies. 11 participants including 1 from space technology laboratory, 2 from a firm of architecture and services, 4 from institutes of computer/information technologies, and others students. They discussed the utility, the feasibility, and the novelty of each created idea, on the oral presentations of inventors with writing on stickers and putting them onto the game board. Also they revised the ideas to fit the requirements of consumers, meaning the users of created service systems in this case.

For example, a created idea highlighted in Figure 2 is *safe landing system for aircraft*, created by combining satellite technology (meant GPS-based monitoring of three dimensional position here) and the hyper air-suspension system controlled by data intensive infrastructure. The GPS based position monitoring was here expected to work for keeping the aircraft at a safe distance from the land in landing, and the hyper air-suspension system to sustain the safety at an exceptional moment when the error of the GPS sensor causes an unexpected contact with the land. As a matter of fact, compressed air has been coupled with oil for the suspension system for aircraft's landing since 1920's, as OPMS (Oleo Pressure Monitoring System), and the data of its inside air/oil pressure and GPS data of the aircraft position are to be integrated for controlling in landing. The integrated data are to be analyzed and processed for optimizing the effect to absorb the impact of collision against the land. As in such a case, we find ideas created in IMG tend to be realized in manufacturing, service industries, power generation etc, although it is not always by the participants of IMG themselves. For more general principles about why and how we should set an environment for innovative collaboration, on which we are improving the IM process, reader is referred to references on collaborative approach for design (Plattner et al 2011, Mohammed and Ringseis 2001, and Gottesdiener 2002).

The Procedure of IMDJ

IMDJ is a specified IM following the three steps below revising the original version in (Ohsawa et al 2013), attended by stakeholders who play as users (consumers in IM) or inventors (data providers, brokers, or analysts). See Figure 3, with reading details in the descriptions below.

***Step 1*)** Data owners can participate in IMDJ without disclosing data, but are encouraged to publish *data jacket*s (DJs hereafter). A DJ is a digest of a dataset, described as meta-data including the names (*not* the values) of variables in a dataset and other potentially

useful information (e.g., summary and expected merit of analysis). A DJ plays the role of a card in IMG – an element for inventing ways to use data.

***Step 2*)** Correlations of DJs are visualized as a game board, for aiding participants in combining DJs. Here, DJs are linked via variables and words in DJs. Causalities between variables and links to other data, if known, may be given by providers of data using RDF, on which arrows expressing the knowledge can be visualized.

***Step 3*)** Inventors create use scenarios, i.e., ideas about how to use data and tools by combining DJs as in Fig.1 and Fig.2. Consumers evaluate, criticize, or buy the scenarios as solutions to their own requirements. As a result, the value of the dataset behind each DJ is evaluated according to its contribution to use scenarios of data. If a participant notices that an essential dataset is missed for realizing a scenario, a new DJ can be proposed.

Let me mention about thoughts behind the three steps above. About Step 1, in the DJ of blood-test data for healthcare, variables may include such blood component as "ASP", "CHE" etc, and the summary can be "data on blood test for faculty members." Such a DJ plays the role of a card of element knowledge in the IMG, in Step 3. On a DJ, the provider may also describe a desirable price of the data, or declare the data is open to pricing negotiations, or that the data is free etc. DJs are shared by stakeholders (e.g., providers, users, potential users, or analysts of data, or data brokers if any) so that they can communicate as in IMG. The width of the society to share a dataset can be also described in the DJ.

On Step 2, let us point out that DJs should not be shown as in a sheer data catalogue where data and metadata are registered and shown in a list (e.g., CKAN), because IMDJ should be an environment where stakeholders discuss the possibility to combine datasets behind DJs, as cards were combined in IMG via discussions. On this board, DJs can be linked via same variables or variables of the same meaning (e.g., "high bilirubin" and "jaundice" that means the same event), or via words common to multiple DJs. Causalities between events corresponding to variables may be declared by providers of data, so that arrows showing the causalities also get visualized and shared intuitively. Providers can also explicitly declare links between data, as in Linked Open Data (LOD: see e.g., Yu 2011). However, this is not required for attending IMDJ, because a point of IMDJ is to discover latent links not easy to declare in advance.

And, for Step 3, the structure of the market of data as visualized in Step 2 enables stakeholders to evaluate each dataset, via discussing how it can be combined with others to create use scenarios i.e., ideas of analysis and the interpretation of the results, toward the discovery of business strategies. As a result, data may be shared free, closed, or sold for an agreed price. The process of thoughts in IMDJ decision can be made on two fundamental principle: First, the visualization of correlations provides hints for combining items connected, via links on the map. Second, the social value of any entity – knowledge, commodities, services, or etc. – can be evaluated via the communication of stakeholders in the market.

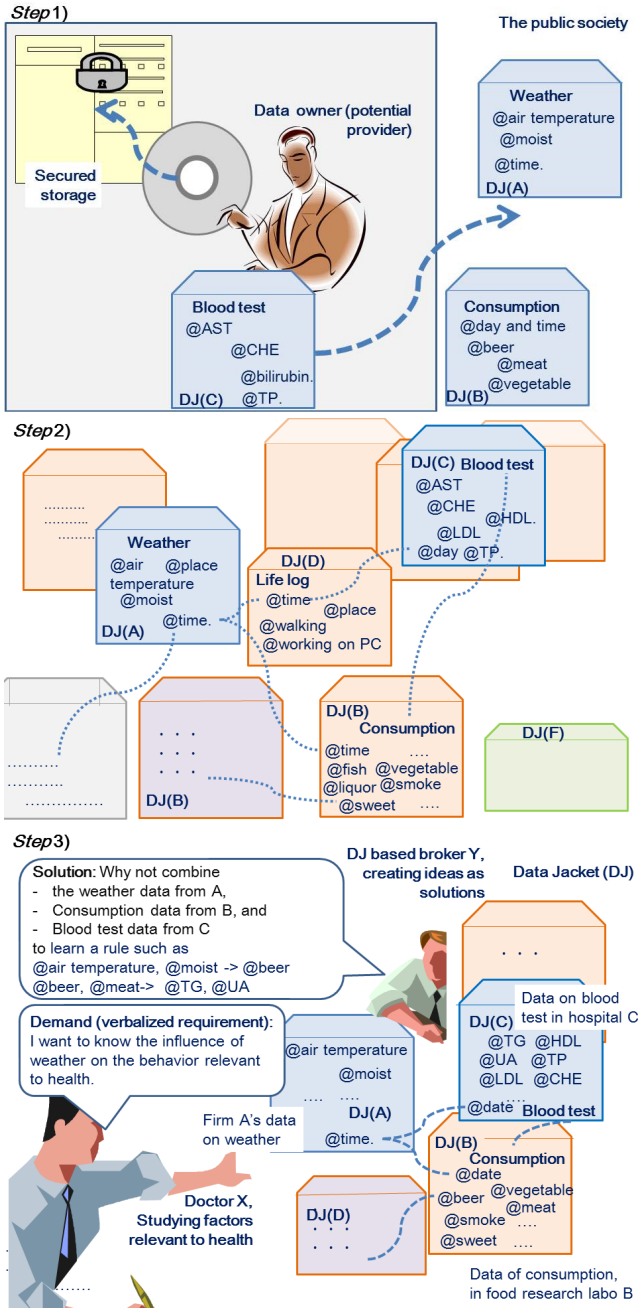

Figure 3 Three steps in the procedure of IMDJ: @X in each DJ means variables named X

Results: Requirements as Qualitative Information of Data Value

In Fig.4, we show an example of IMDJ with 23 DJs. In this example, we find personal data represented by DJs such as:

DJ6) Users' profiles and comments in Facebook: @name, @affiliation, @interests, @friends, and

DJ16) Economic (& health) data of countries: @GDP, @consumers price index, @deaths by cause, @health systems performance,

are combined to create

Solution 2 (DJ16 & DJ6): A medical care service on such patterns (association rules with confidence and support) as:

@interest -> @disease;

@friends, @affiliation, @commenting frequency -> @health state;

In this manner, we find qualitative data (DJ6) and quantitative data (part of DJ16) can be combined to create solutions for given requirements. In this case, the idea above was created responding to the requirement to take advantage of data for medical care services. As a result, we find each stakeholder may find the value of others' personal and qualitative data, which are not usually regarded as essential for one's own heath or life.

For 14 cases of IMDJ, where 15 to 40 DJs has been used in the initial state of the game, we counted $n\_{solution}$: the number of presented ideas, called solutions in this section, $n\_{requirement}$: the number of cards describing requirements, n_{DJ}: the number of DJs prepared in advance, n_{newDJ}: the number of DJs added in the game. As a result, as in Fig.3, created ideas and new DJs proposed in IMDJ tend to increase significantly with the number of requirements. Here the Pearson correlation coefficient r is equal to 0.70 between $n_{solution}$ and $n_{requirement}$, 0.23 between n_{newDJ} and $n_{requirement}$. In comparison, r = 0.43 for $n_{solution}$ and n_{DJ}. These all take positive values, but the significant dependency between $n_{solution}$ and $n_{requirement}$ is noteworthy because requirements counted by $n_{requirement}$ mean the pressure to restrict the search space of solutions, whereas the prepared elements counted by n_{DJ} mean the resource for enlarging the search space. This maybe linked to the discovery in cognitive science (Finke, Ward, and Smith 1996), in that constraints rather reinforce creativity than suppress it – although Finke measured the quality instead of quantity of created ideas. We may focus on another interesting feature of Fig. 5, i.e., the peak of n_{newDJ} near $n_{requirement}$ of 30. The interpretation of this feature will be discussed in our forthcoming presentations, but the positive trend for $n_{requirement}$ less 30 is noteworthy (r=0.73), in that we can restrict the number of requirements to less than 30 in order to deepen discussion for each requirement so that participants can externalize the necessity of additional data.

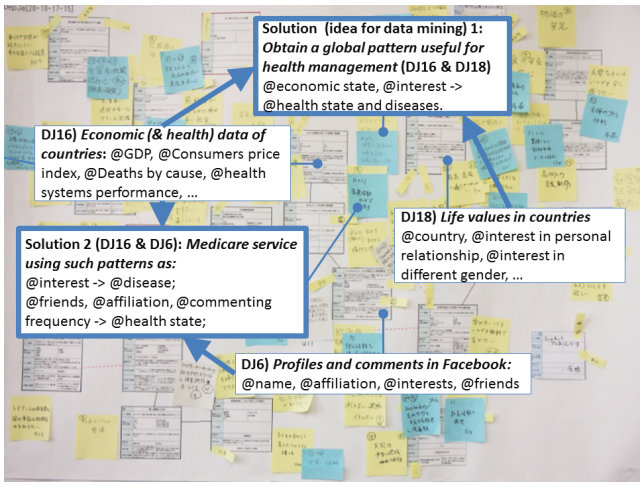

Figure 4 A result of IMDJ: Datasets combined to create a method for medical care systems

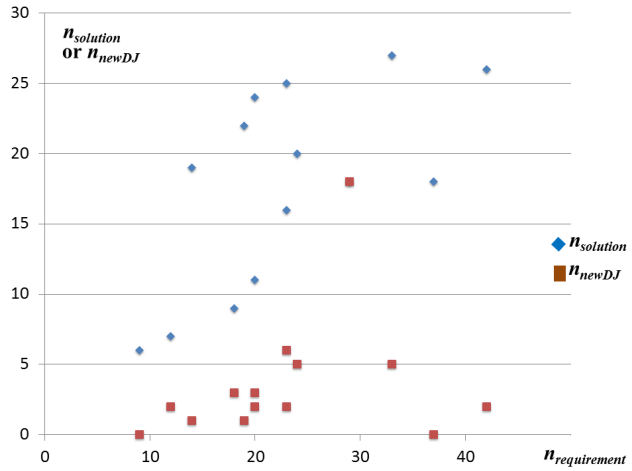

Figure 5 Correlation of the number of requirements versus proposed solutions (r=0.70) and versus proposed new DJ's (r=0.23).

Conclusions

As in the example shown in the section of experiment, a point of IMDJ is that participants are enabled to discover qualitative meanings of existing datasets, via communication on DJs that are the digests of data. The qualitative meanings here are given as the requirement for combination of datasets, i.e., the reason why a consumer requires inventors to present an idea for combining existing datasets, and as the reasoning of inventors about how to satisfy the requirement. Such a requirement may first be given by consumer's putting the game board, or by the consumer's externalizing the requirement on hearing inventors' presentation of ideas that can be linked to latent requirements.

In this sense, the second item in the listed challenges of this workshop (Kido and Takadama 2014), i.e., "(2) how to turn the large volumes of impersonal quantitative data into qualitative information that can impact the quality of life of the individual" is to be realized here by discussing the requirements to be satisfied by a set of DJs. This item is also contributive to the first and the third, i.e., "(1) how to quantify our health, wellness, and well-being for generating useful big data that will become meaningful knowledge" and "(3) how the quantitative data and qualitative information contribute to improving our health, wellness, and well-being" by using IMDJ. For (1), inventors propose to add new DJs in the communication as stated in Step 3 of IMDJ in the previous section, which calls for the generation of new data if they do not exit yet. And, for (3), DJs ranging over quantitative and qualitative datasets are combined in the brains and communications of participants in IMDJ.

Thus, the market-wise communication can be expected to be a key approach toward externalizing and combining the values and use scenarios of datasets in real application domains. Note we are calling potential participants of IMDJ for DJs, to which reader can respond from the entry form as in: http://www.panda.sys.t.u-tokyo.ac.jp/ .

Acknowledgments

This study has been essentially supported by JST CREST, and some experiments were aided by the Nuclear Regulation Authority Japan. We also appreciate participants and supporters representing the private section of Japanese industry.

References

Kido, T., and Takadama, K. 2014. AAAI Spring Symposium on Big Data Becomes Personal: Knowledge into Meaning — For Better Health, Wellness and Well-Being (scope of this workshop).

Bollegala,D., Kushimoto,M., Yoshida,Y., and Kawarabayashi, K. 2013. Mining for Analogous Tuples from an Entity-Relation Graph, in Proc. International Joint Conference on Artificial Intelligence.

CKAN. http://ckan.org/

Dervin, B. 1983. An overview of sense-making research: concepts , methods and results. Presented at the Annual Meeting of the International Communication Association. Dallas, TX.

Finke, RA., Ward, TB., and Smith, SM. 1996. Creative Cognition: Theory, Research, and Applications, MIT Press.

Fruchter, R., Ohsawa, Y., and Matsumura, N., 2005. Knowledge Reuse through Chance Discovery from an Enterprise Design-Build Enterprise Data Store, New Mathematics and Natural Computation 1 (3): 393—406.

Gentner, D. 1983. Structure-mapping: A theoretical framework for analogy. Cognitive Science 7: 155-170.

Hong, C.F., Yang, H.F.,Lin, M.H., Lin, G.S. 2006. Creative Design by Bipartite KeyGraph based Interactive Evolutionary Computation, Lecture Notes in Computer Science 4253, Springer, 46-56.

Joshi, M.V. 2002. On Evaluating Performance of Classifiers for Rare Classes. Proceedings of The 2th IEEE International Conference on Data Mining (ICDM 2002) 641-644.

KDnuggets : Datasets for Data Mining and Data Science, http://www.kdnuggets.com/datasets/index.html.

Llor, X., Goldberg, D.E. et al. 2006. Innovation and Creativity support via Chance Discovery, Genetic Algorithms, New Mathematics and Natural Computation, 2 (1): 85-100.

Ohsawa, Y. 2005. Data Crystallization: Chance Discovery Extended for Dealing with Unobservable Events, New Mathematics and Natural Computation 1 (3): 373 – 392.

Ohsawa, Y., Matsumura, N., Okazaki, N., Saiura, A, and Fujie, H. 2006. Mining Scenarios for Hepatitis B and C, Paton, R. (ed), Multidisciplinary Approaches to Theory in Medicine, Elsevier, 209-231.

Ohsawa, Y., Kido, H., Hayashi, T., and Liu, C. 2013. Data Jackets for Synthesizing Values in the Market of Data, Procedia Computer Science 22: 709-716.

Ohsawa, Y. and McBurney, P. 2003. Chance Discovery, Springer.

Ohsawa, Y., and Nishihara, Y. 2012. Innovators' Marketplace: Using Games to Activate and Train Innovators, Springer.

Ohsawa,Y., Benson, N.E., and Yachida,M. 1998. KeyGraph: Automatic Indexing by Co-occurrence Graph based on Building Construction Metaphor, Proc. Advanced Digital Library Conference (IEEE ADL'98): 12-18.

Weiss, G.M., & Hirsh, H. 1998. Learning to Predict rare Events in Event Sequences, Proc. International Conf. KDD'98, 359-363.

Yu, L. 2011. A Linked Open Data, Developer's Guide to the Semantic Web Chapter 11, Springer: 409-466.

Plattner, H., Meinel, C., and Leifer, L., eds. 2011. Design thinking - Understand - Improve - Apply -, Springer.

Mohammed, S, and Ringseis.E. 2001. Cognitive Diversity and Consensus in Group Decision Making, Organizational Behavior and Human Decision Processes 85(2):310–335.

Gottesdiener, ,E. 2002. Requirements by Collaboration: Workshops for Defining Needs. Addison-Wesley Professional.

Windows Azure Marketplace. https://datamarket.azure.com/.