2nd Technical Committee Meeting　 Summary of Minutes

Date & Time：Jan. 9, 2013 (Wed.) 13:30～15:30

Venue：MRI, 4th Floor, CR-C

Participants (Honorifics Omitted)：

Main Researcher：Noboru Koshizuka (Professor, Interfaculty Initiative in Information Studies, the University of Tokyo)

Sub Researcher：Hideaki Takeda (Professor, National Institute of Information)

Committee Members：Kenji Hiramoto (Executive Advisor for Chief Information Officer, Ministry of Economy, Trade and Industry), Yoshiaki Fukami (Visiting Researcher, Keio Research Institute at SFC), Akihiro Nakao (Associate Professor, Interfaculty Initiative in Information Studies, the University of Tokyo)

Observers：Information and Communications Bureau of Ministry of Communications, IT Strategy Office of Cabinet Secretariat, Commerce & Information Policy Bureau of METI, Geospatial Information Agency of MLIT

Others：Katsunori Shindo (Yokosuka Research Park)

Secretariat：Takeshi Tsuguni (MRI)

Documents Distributed：

2nd Technical Committee Meeting Agenda

Ref. 2-1　Seating Chart, 2nd Technical Committee Meeting

Ref. 2-2　List of Committee Members

Ref. 2-3　Trend of Vocabularies as the Base of Open Data in Major Developed Countries

Ref. 2-4　Important Points for the Discussion of the Technical Committee

Proceedings：

1. Opening & Confirmation of Distributed Documents
2. Self-Introduction of Mr. Nakao, Committee Member

* I teach computer science application in the Comprehensive Information Analysis Course of the Interfaculty Initiative in Information Studies, the University of Tokyo, which happens to be the same Initiative as that of the senior researcher of the Committee. My major is Information Network and I am making research on the Internet. I’ll be happy if I could contribute to Open Data from the viewpoint of data available from the information network.

1. Important Points for the Discussion of the Technical Committee

* We have two major items to discuss today.

First, we ask the Secretariat to explain what they have organized in terms of planned

form of output, and we would like to get your basic approval of what was proposed by

the Secretariat. Second, we ask for the information sharing from METI which has been

investigating cases of the NIEM of the US and ISA of Europe. These are essential items

when we talk about approaches of developed countries for making public / government

information open, and quite useful for the Technical Committee to reflect them to our

own technical standards and to keep consistency between theirs and ours.

The circumstances surrounding open data have been changing since the establishment

of the Consortium. IT Strategy Headquarters has also getting active. It is desirable if

the Consortium could make some contributions in the circumstances. Therefore, today’s

discussion may be a little expanded in the scope. Taking account of various views

expressed in the previous meeting, I hope we can come to an agreement on the direction

of our opinion integration of this year.

Initially, we considered brushing up the standard API and vocabularies provided by the Infrastructure Development for Collaboration for Data Circulation administered by the Ministry of Communications for use by those who wish to disclose their data. However, we noticed the existence of different stages and players of open data. Therefore, we may have to consider approaches not only to high-end users but also non-high-end users.

In other words, we may be required to consider a coherent system as a whole which duly accommodate specifications of API and technical matters on one hand, and an intermediate system for non-serious users of open data together with easy-to-understand guidebooks etc. on the other hand.

As for the Ref. 2-4, since it is described significantly in detail, we may need a stepwise

explanation broken down to 2 or 3 steps. I would like to ask the Secretariat to explain it

accordingly.

・Mr. Shindo explained from page 1 to 5 of the “Ref. 2-4 Important Points for the Discussion of the Technical Committee”.

* Do you have any questions so far?

In connection with the disclosure of data, one of the frequently-asked-questions is if CSV format is OK or not, and how can they disclose in CSV. As mentioned in “Desirable Standards of Data and API” of Point 4 of the Page 4, most of the statistical data of the government’s White Papers, for example, are only in PDF format. It is natural for the government to wonder that if they have to release them in the framework of open data, what they should do exactly.

The aim of Points 4 and 5 is that the Committee should cope with the issue of the preparation of something like guidebooks to explain the usage of open data based on the concrete data. This should be undertaken in the course of our step by step action toward the realization of ideal open data. Not only limited to CSV, but we should also look into any other systems that are strongly requested by data owners or users.

Compared to the discussion in the previous meeting, the importance of the IT Strategy Headquarters as one of the institutions we have to bear in mind increased. Since the Technical Committee is responsible for technical matters, it is desirable if we can produce output for them.

Next, we would like to discuss data standards and API standards for open data. Generally speaking, we need a macro perspective through which we look down and understand what type of data are suitable for open data and what type of data standards are applicable for open data. Upon reviewing the whole picture of these, as we do in taxonomy, we should propose the way of optimum usage of open data in accordance with different stages. I ask the Secretariat to explain it in this context.

・Mr. Shindo explained from page 6 to 56 of the “Ref. 2-4 Important Points for the Discussion of the Technical Committee”.

(Comments and Q & A regarding the Ref. 2-4)

* Thank you for the explanation of the targeted data for open data that our Committee has to deal with, and outlines of their existing data standards.

In terms of API standards, there have been changes of popularity, and currently there are various types all of which can be categorized by their respective characteristics.

Our basic idea is to choose and adopt appropriate standards commensurate to data characteristics at the time of making data open. The categorization indicates that we do not need to create new data formats or protocols from the scratch. However, even in the case of the selection of the optimum standards, data users must decide vocabularies, IDs, ways of dealing with Chinese characters in case of Japan, ways of preparation of data catalogues and meta-data, etc. by themselves to some extent. The Technical Committee may need to discuss the matters as well.

As for API standards, while what we must decide is not many, items for consideration for the actual API standard application are well organized in the page 54. In addition, we need to consider the trends of existing systems, consistency and interoperability between theirs and ours.

I invite comments or opinions over things discussed so far.

* In page 6, with respect to the overall image of open data for which the Committee is responsible, statistics, geographical data, sensor data, goods distribution and disaster prevention etc. were exemplified, in the previous meeting, as data the Committee should target. This time, while disaster prevention dropped, administrative information data significantly increased. Is this because of changes of the circumstances?
* Targets have not changed. We just increased examples.
* This description gives an impression as if the Committee will handle all these. I suggest to describe that the Committee will intensively deal with the prioritized items. Later, we are introducing the METI’s approach to you. As some of your described items overlap those handled by the METI’s approach, I think it is necessary for you to clarify the targeted data.
* The meaning of handling is diverse. Here, we included those we directly handle and those existing data which we are going to adopt.
* As we make research with limited resources, it is better to clarify the data to adopt from other sources, and the data the Committee handles and the person responsible for them.
* Based on the list of items, we may add various information and consider the ways of response to them, and hope them to be developed into a map in future.
* You may think about adding implication of involvement of METI to page 3, for example. In terms of administrative data, it is better accepted by the public if the cooperation of METI is clearly mentioned. And you can save unnecessary labor.

Next, at page 10, the categorization of “data” and “data-like contents” seems ambiguous. For example, I am curious what you are going to categorize twitters that are data-like text data. Data or data-like contents?

* In open data, I think data means encoded information that is machine-readable. The disclosure of documents is an important factor in open data. Documents are surely in the category of contents. I make it clear that documents are included in our efforts of open data.
* I wonder if the upper row is for the generic information, while the lower row is for contents including data. Since we are not going to deal with all kinds of data, it might be OK to define as “digital information including data” if we are to clarify the scope of our work.
* Problems are expression and scope. In terms of scope, there is a risk that data might be erroneously integrated into graphable numerics rather than codes. This may be welcome for those who, including myself, are for wider scope of data. But we have to think about expression. In addition, from where we are going to expand the scope of data, and how we prioritize data are other issues. If we agree to the major premise that data useful for public interests should be made open as much as possible, not in the form of mere numerics or codes, but in the form that is easy to use, the next thing to do is to discuss what procedure is needed and what we should appeal to the public.
* At the time of decision of the scope, I presumed the inclusion of the high-priority data to hasten the consideration process. However, after reviewing the movement of Infrastructure Development for Collaboration for Data Circulation by the Ministry of Communications as well as IT Strategy Headquarters, I noticed significant varieties of high-priority data. While there is CSV which is comparatively easy format such as statistical data on one hand, there also exist documents such as White Papers on the other hand. They are both extremes, but they are both highly needed information.

I think both of them should be included in the scope, and be treated with high priority.

* In Japan, the definition of information disclosure and its real activities are confused. This should have been rectified earlier in the stream of information disclosure initiated by the government. Currently, ordinary definition of open data mistakenly includes the mere information disclosure. While this issue might be handled by the Open Data Working Level Meeting, we may also have to think about the clear separation of open data and information disclosure or just keep them unseparated.

In the past, there have been cases of inappropriate disclosure of documents. This time, we may think about clear inclusion of documents as data in the scope.

* I have been repeating to clearly say that the difference between information disclosure and open data is that open data is to provide data in the format that is machine-readable and program-writable. However, as many other people told me that was too much, and I had to compromise to some extent.
* Inadequacies in the past can be overcome by inclusion of information disclosure into open data.

The first thing we must do is to settle the license issue. As the result, the usage of White Papers will become clearer in addition to the license issue settlement. In that sense, inclusion of documents as data in the scope is not bad.

* Whenever I come across with document files while I am preparing some materials, I strongly feel the needs of inclusion of documents as open data.
* At the page 9, in terms of static and aggregated data versus real-time data, the proposed categorization in accordance with the difference of data collection method, access method and implementation makes sense. On the contrary, I do not understand the categorization of the page 10. My view is the simple indication of inclusion of data such as White Papers is enough and better than categorization.
* This is a matter of urging positive attitude rather than categorization. If I put it in an extreme terms, copying documents with pencil at government office after the strenuous efforts of requests for disclosure based on the ordinance followed by approval by court’s order or judicial precedents, etc. is also information disclosure. On the contrary, positive disclosure of data for the benefit of users is a matter of attitude, not the matter of technology. What we must do is to do the utmost to the technical affairs that support the data owners’ positive attitude.
* I have concerns on 2 points in the International Standardization, at page 3.

I wonder if there are any precedents, in foreign countries, of registration of API standards etc. to ITU-T. In a big framework, there exists W3C for RDF.

If there exist any advanced cases with regard to open data, I am concerned about the

consistency between theirs and ours. If not, I am curious if we had better quickly move for international standard setting.

Another concern is what we are going to standardize when we talk about international standardization. It is easy to understand that we need to standardize new data format and API standards. But, if we adopt something from arrays of existing formats or standards, are we going to propose the ways of usage, or to propose the new system? I am not sure about the output of our work.

The creation of new guidelines for proposing to the Open Data Working Level Meeting is one of outputs easy to understand. However, it is not clear what are already existing for standardization and what the Committee is going to newly standardize.

* This year, we are going to survey which institutions are doing what sort of activities. While I do not have enough knowledge about static and integrated data, as for standardization of real-time data, I do not think any institutions have registered their standards to ITU-T. As for W3C, since I do not have the whole picture of it, I have no idea about its details.

In terms of static and integrated data, ITU-T seems to have done something regarding vocabulary standardization of contents. However, since they are majoring in multi-media, their activities are quite different from what we are doing now. The current hot issues are how to integrate data collected from real-time data sources or from sensors into the cloud computing, and how to integrate sensor data protocols at the time of handling by private sector telecommunication business servicers, etc.

* Are they doing in the so-called smart grid as well?
* Various institutions including ITU, ISO and IEC, etc. have been considering how to collect sensor data from wide range of nets, and extremely wide variety of issues are emerging.
* A part of the W3C standardization was approved as the fast track version by ISO. However, it took four to five years to be ISO standards after they were recommended by W3C. It also took years for HTML4.01 to be JIS standard after the approval by ISO.

As W3C declared themselves to be a de facto organization as well as an industrial consortium, they have no enforceable power. The Web standard is basically de facto, and as everyone use it, others also use it. W3C is in reality an integrated organization for patent-free usage. Basic items including RDF, Mark and Dublin Core, etc. are used in communities. Although W3C is also a community, inter-community liaison or consistency is an important issue for them. While W3C, IETF and ISOC are de facto organizations, their activities are of public nature to some extent. They themselves are conscious of the issue of the liaison and consistency with such organizations as ISO, ITU and IEEE, etc.

* In terms of the output of standardization, if we use the existing data format, the core concern of standardization will be vocabularies. Who is responsible for standardization is another issue. For example, talking about the vocabularies used in the food traceability data that has been administered by the Ministry of Communications, there exist no international standards for such vocabularies. In fact, the number of data with vocabularies are limited. In other words, 4 issues mentioned in the document have the possibility of standardization.

Another issue is how we deal with government data and how we describe human data.

In future, in case the numbering system for the whole Japanese population is in place, it is ambiguous if the human data and the model to be adopted are compatible or not. If it is not compatible, the nature of human numbering system should be changed. Therefore, in terms of vocabularies, there will be much room for standardization.

Talking about protocols, REST is still in the model level, and we have to decide actual protocols one by one. Aside from the merit or demerit of standardization, there exists a possibility of it.

As the trend of ISO, the standardization of procedures for open data and their guidelines etc. is on the international main stream. Following the trend is important.

As for the ways of coordination between meta data and mutual operation, there still exist needs of standardization.

The standardization between sensor networks has never been completed yet.

Unless the standardization between machines is realized, the network does not function.

Should the government take initiatives for these matters or not? Is it easy for the government or not? There are so many things to think about.

* With respect to API, here in the document, only examples of data access are listed. We need to elaborate on what sort of data retrieval can be made.
* When we talk about standardization, it is not the classical standardization. For the first place, data exchange must be available in open data and LOD etc. without any classical standardization. For realization of data exchange, some agreement may be needed.

Currently, as organizations for standardization are also increasingly active in this area, there may be possibility of international collaboration for standardization. I hope to discuss this matter in the Committee on how we should react to the movement.

* Coming back to the topic of HTML4.01 and ISO, the standardization of W3C recommendation was made in Dec., 1999, and the first file of HTML4.01 of ISO version was released on Mar. 2000. However, as members had to bring it back to their home countries for trial before the feedback, the final version was released in 2003. It took three years for actual examination in home countries. As the research on HTML4.01 was initiated in the mid-1990s, it took nearly 10 years up until the inauguration of its practical use. It is critically important to expedite the trial implementation prior to the stage of dejure standard so that comments can be submitted and reflected for the ISO standardization. We should think that the trial implementation precedes the standardization.
* If it is written as ITU-T, we may tend to think that it is the classical type of standardization which should go through the dejure standardization process, but it may be de facto standard in reality.

This is an essential topic in case of M2M or smart grid etc. that communicate between sensor-type machines. We should pay attention to their movement.

* When we think about output of M2M, it should be noted that we must consider not only the protocol for data exchange at low level, but also the high level automatic exchange of data semantics. Although the vocabulary decision is necessary at the highest level, the sensor network has not yet reached that level as discussions on network wares and transport, etc. are still going on. However, they are conscious of the needs when they talk about the general framework architecture. It might be a good idea to provide them with this sort of input.
* We are ready to provide any proposals. However, we should be selective in pushing forward to standardization. Irrespective of dejure or de facto standards, we should assess which one is prospective to take large shares in future. The standardization is not the final destination. We must ensure the future usage of our standards. Otherwise, even if we prepare standards, it is useless if the standards are not put into use or out of fashion in the world arena.
* I would like to raise another important factor. As the government is expected to be a major player in open data, the open data system will be procured through the public bidding, which means the necessity of consideration of the relationship with WTO. If we have any examples of domestic open data, any systems relevant to them may be given high priority in this regard.
* With respect to character codes of Issue 3 of page 43, I do not think it is necessary to include it in Issues. Although it says “Parts impossible to describe in Unicode exist.”, it is now possible to describe even the characters of family register. In a few years, it will be possible to describe everything in Unicode. As the Ministry of Justice says external characters not used in family register are not allowed to use in their documents, this issue has been virtually solved. In parallel with Unicode, there are descriptions of three letters of CHN. As this can be solved by putting JA or CN in sentences, we may not need to think about it too much.
* We do not intend to examine it here. We just simply put it as reference. It may have been better not to include it in Issues.

1. Trend of Vocabularies as the Base of Open Data in Major Developed Countries

・Mr. Shindo explained Ref. 2-3 “Trend of Vocabularies as the Base of Open Data in Major

Developed Countries”.

(Comments and Q & A regarding the Ref. 2-3)

* In around every November, we have TPAC General Meeting where we have a discussion track to talk about open government. Topics discussed there regarding open data are roughly divided into two. One is integrating use cases, and their discussion has been generally going along vocabularies and data sets. The other one is purely technical, and their discussions have been such as how to deal with JSON in addition to XML-RDF, how to expand the open data in the framework of Five Star, and how to well prepare LOD applications, etc. Their characters are different, and they have been up and running in parallel. There was a report that in European countries they have been integrating cases and proposals in terms of use cases. Basically their movement is limited to use cases. The core technology simply links to triplex and RDF/XML. Changes have been taken place toward standardization of appropriate vocabularies and application creation know-hows based on use cases. Therefore, it is important to quickly create practical use cases employing existing materials, and provide knowledge earned from that experience rather than considering everything from the scratch.
* While UCore is disappearing, UCore has developed a set of applications. We may be able to utilize their knowhow. In connection with W3C, an international conference call is going to be held on coming 21th of January. This conference call is the venue to share the approaches of open data of member countries. This time, it is the turn for Japan to introduce its approaches, though the presentation will be given by the US consultant on behalf of Japan.
* NIEM and ISA started as data exchange format between government offices in order to realize the E-Government, and therefore disclosing government data was not necessarily the primary objective. However, my understanding is it is OK to disclose those exchanged data when open data is realized. Is this correct?

I want to ask METI if its approach is NIEM as a tool or NIEM as an objective.

* What we are aiming at is not for objective, but creation of tools.
* Does it mean METI is heading for Open METI making use of NIEM framework?
* If data items increase, it will contribute to open data, and may lead to IEPD-type data.
* I think it is a right way not to jump into open data all of a sudden, but disclose data based on the E-Government system where inter-government data exchange is undertaken electronically. In this context, I am impressed by the US way of doing.

1. Draft Specifications to be Prepared by the Technical Committee

・Mr. Shindo explained Ref. 2-4 “Important Points for the Discussion of the Technical Committee”, from page 55 to 74.

(Comments and Q & A regarding the Ref. 2-4)

* Draft Specifications were explained in the previous meeting. This time, it is wider in scope. As mentioned in page 56, we suggest the Consortium produce various types of technical guidebooks as its output, and give them as input to the E-Government Open Data Working Level Meeting.

As for CSV standard mentioned in page 72 and 73, although it is still in the stage of just an idea, the provision of something like a guideline indicating that positioning of RDF as the supreme standard will greatly help etc. may make sense. We hope that the Technical Committee would produce not only RRDF-based Draft Specifications but also an easy-to-understand guideline that simply treats table-type information as they are.

Do you have any questions so far?

* At page 56, there is a description “Recommendation Guide for Open Data of Public Documents” in the column of Technical Guide. I think this is related to the work of the Governance Committee. The issue of license, for example, is one of them. I wonder if this is prepared as the Technical Guide or prepared by the Consortium as a whole. If it is mentioned simply as the table format, it is easy to understand. But I do not think the simplification to that extent is not possible.
* As a guide, it may have to include the social process as well. In terms of technical matters, not only the Technical Committee, but also the IT Strategy Headquarters is responsible. However, as for the process, their situations are different from ours, and we are not sure if our inputs are applicable to theirs. The issue of license, etc. is fairly influenced by the users, and I think it may be a deviation from our scope of work if we input something which was discussed and prepared by the Governance Committee. It may be better to change the title of “Recommendation Guide for Open Data of Public Documents” and focus on purely technical matters.
* Yes, I agree. We should do so.
* Although we cannot name specific licenses, we can mention technical procedures of open data including examples of licenses, selection of necessary licenses, examples of popular usage, etc. In order to give consistent guidance to officers in charge of open data, we include these information.
* There are the Technical Committee and the Governance Committee in the Consortium. Similarly there are the Data WG and the Dissemination WG in the E-Government Open Data Working Level Meeting. The Governance Committee is also currently discussing how they can prepare something useful for officers in charge of open data. We may be able to contribute something for them.
* I understand outputs from different sources will be finally merged. However, the decision of the responsible institution for the finalization will definitely need the coordination.
* At the risk of overstepping the border, we should actively go farther for producing outputs. As this is a highly fluid area, if everyone is too shy and modest, there will surely be things unattended by anyone in the end.
* In case of any more comments and views, I request you to send them to the Secretariat by e-mail.

1. Schedule of the Next Meetings

* As for the timing of the 3rd and 4th Meeting, your cooperation will be appreciated as we coordinate it by the e-mail exchange.

End