第 2 回 SPARC Japan セミナー2015
「科学的研究プロセスと研究環境の新たなパラダイムに向けて
－e-サイエンス，研究データ共有，そして研究データ基盤 -」

「研究のバリア」を打破する 研究基盤デザインと研究データ利活用

北本　朝展
（国立情報学研究所）

講演要旨

現在の学術研究には様々なバリア（障壁）が存在しており、それを打破することは学術研究の可能性を広げるために重要な課題である。コミュニティを隔てるバリア、データの活用を阻むバリアなど、様々な形で存在するバリアをどのように乗り越えていくか、そのための技術と戦略が問われている。そこで本発表は、発表者自身がこれまで構築してきた「デジタル台風」「ディジタル・シルクロード」「東日本大震災デジタルアーカイブ」などの研究基盤を事例とし、研究基盤のデザインや研究データの活用においてどのようなバリアに直面し、それをどう打破してきたかを紹介する。また、こうしたバリアへの対応は「オープンサイエンス」においても重要な課題となっており、その解決に向けて情報学が貢献できることについても私見を述べる。

北本　朝展

国立情報学研究所 コンテンツ科学研究系 准教授。東京大学工学系研究科電子工学専攻修了。博士（工学）。大規模な実世界データから価値を創出する研究に興味を持ち、地球環境データや災害データから人文科学データまで、異種かつ多分野のデータを対象とした統合・解析・可視化等の方法論に取り組む。最近はオープンサイエンスを取り巻く活動にもかかわり、他分野の研究者や専門家と協働しながら、研究基盤を中核とした学術研究の新しい展開を目指す。文化庁メディア芸術祭アート部門審査委員会推薦作品など受賞。

今日は、研究のバリアをどうやって取り除いて新しい研究をしていくかという話をします。バリアはいろいろなところにあります。例えば、あなたと他者の間にあります。また、機関の間にも、研究コミュニティーの間にも、研究者と市民の間にもバリアがあります。こういったものに対してどういうことをやってきたか、実例のご紹介をして、ソリューションについて考えたところをお話しします。

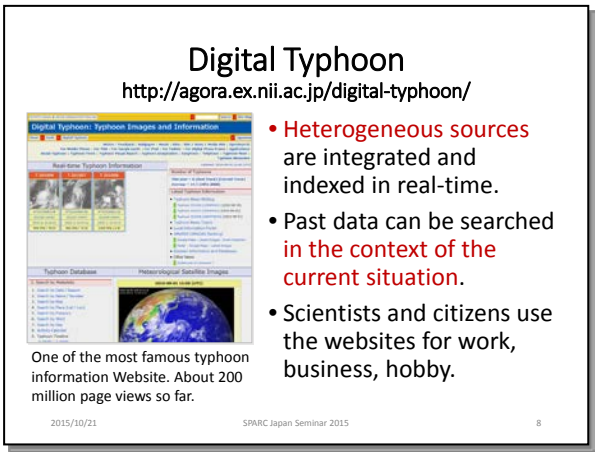
私の定義としては、オープンサイエンスとは、今の研究のやり方に不満を持っており、オープンという言葉が研究を変えるキーワードであると考える人たちの、さまざまな考え方が集まる場所です。オープンサイエンスは、それぞれの人が違うことを言っているけれど、希望が集まる場所は 1 カ所だという感じで捉えていただければと思います。そのようなオープンサイエンスの動きの中で、研究データをオープンすることでどのようなメリットがあるかということが議論になります。そこで、私が行っている二つの研究事例をお話ししたいと思います。

デジタル台風

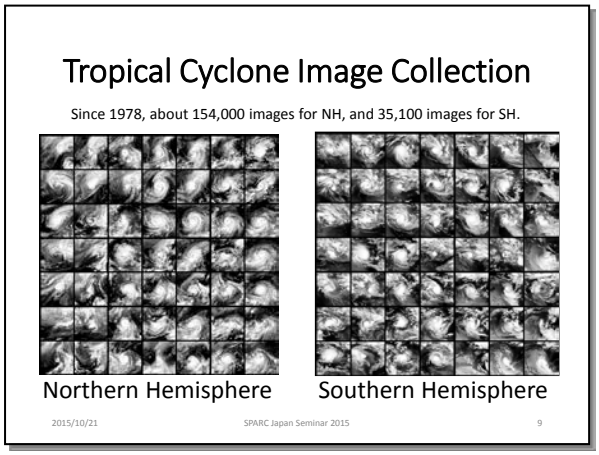
最初は、「デジタル台風」という、主に自然科学データの話です。デジタル台風というのは、現在発生している台風の情報が出ており、その裏に過去の莫大なデータが入っているデータベースです（図 1）。今日は台風が二つ発生しているので、二つ情報が出ています。これはいろいろな情報を集めて、統合して検索できるようにしています。現在の状況に似たような過去の情報があるかも検索できます。これは研究者だけではなく、市民などいろいろな人が使っています。意外な利用方法があり、例えば、あるサーファーはこのサイトを使って、次にどんな波が来るかをチェックしています。最近もサーファー雑誌から掲載許諾依頼が来ました。サーファーはこのデータベースを使って、「過去の伝説の台風のときはこんな波だった」ということを語り合う人が多いのです。今まで約 2 億ページビューと、非常に多くの人が使っています。

コアになるデータは図 2 のようなもので、気象衛星「ひまわり」から台風の部分だけを取り出した画像のコレクションをつくります。全体で約 18 万件、このような画像をひたすら集めていきます。そうすることによって、画像の検索をしたり、検索技術の研究に使ったりできるのです。やっていることは、データベースに過去のたくさんのデータが登録されているので、現在来ている台風に似たような過去の台風があったのか、いろいろな観点から検索するということです。例えば、雲の形が似ている台風があったか、ニュース記事に書かれている災害と同じような災害が起きた台風はあったか、今すごい雨が降っているが、これと似た雨のパターンの台風はあったかということがリアルタイムで検索できるようにしています。例えば経路データでは、過去に似た経路の台風があったか、雲パターンのデータでは、過去に似た雲パターンの台風があったか、雨パターンのデータでは、過去に似たような雨パターンの台風があったかなどがリアルタイムで検索できます（図 3・図 4・図 5）。このように、いろいろな種類のデータを使って分析するというのがこのサイトがしていることです。

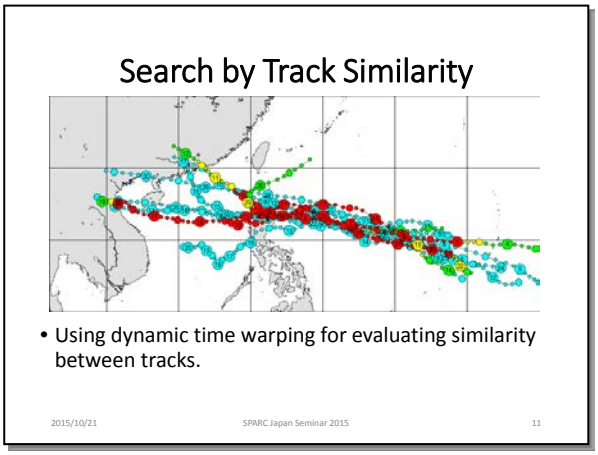
さらに、センサーだけではなく、人間の集めるデータも使えます。例えば、雨が降ってきたというツイートと、雨が降っているというレーダーのデータを組み合わせれば、人々が感じる雨はどこに降っているかということが分かります。これはそう簡単ではなく、ツ



（図 1）



（図 2）



（図 3）

イートから地名を抜き出して、地名を緯度経度に変換して、それをマップしています（図 6）。そのために、GeoNLP というソフトウエアをつくり、文章を分析して、その結果が間違っていたら地名辞書を増やすというフレームワークをつくっています（図 7）。

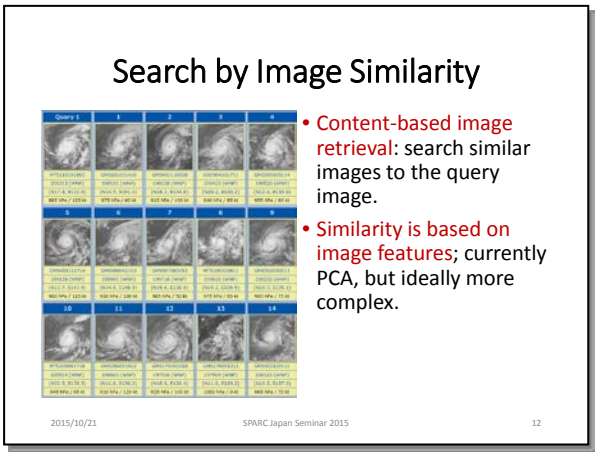
その一例として、例えばどこで雨が降って、どこで雪が降っているかということをツイートから分析すると、雨が降っているところは青、雪が降っているところは白でマッピングされ、東京で雨と雪の境界がどこにあるかということが分かります（図 8）。

このようにいろいろなデータソースがあり、例えば、先日の台風災害でもそうでしたが、河川の情報と雨量の情報が別々の場所から出てくるので一緒に見られないという問題があります。自然としては分かれていないのに、観測する機関が分かれているというバリアがあり、統合できないのです。

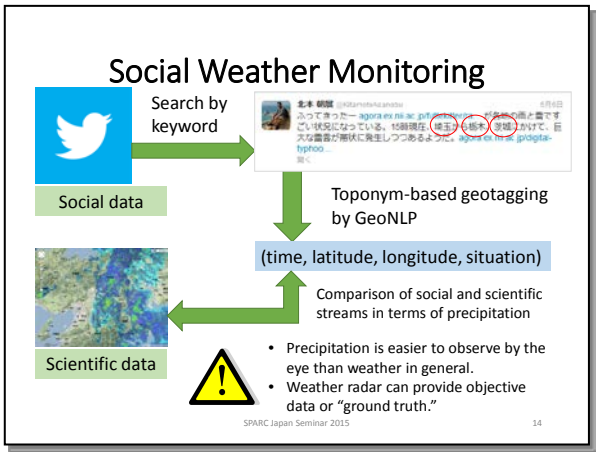
このインフラで心掛けているのはバリアフリーです。このようなリサーチデータのインフラをつくるときに、研究者用と市民用、プロフェッショナル用とアウトリーチ用というように分けられることがよくありますが、私は両方が使えるようなユニバーサルなインフラがいいと思っています。それは、結局、市民の使い勝手を良くすると、研究者の使い勝手も良くなるからです。いろいろな人が使えるように使い勝手を良くすることも、バリアフリーということでバリアに関係するところです。

Digital Silk Road

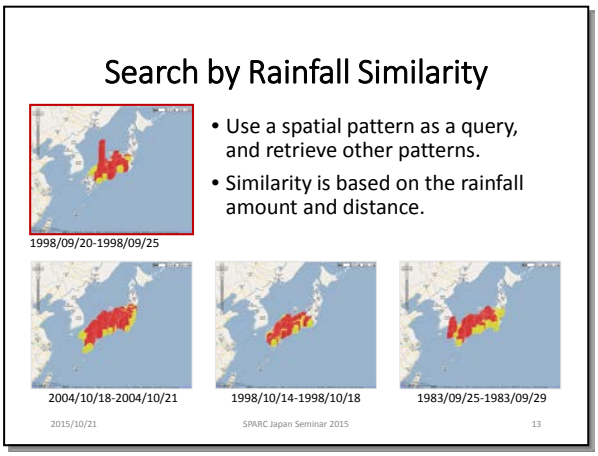
次は全く毛色が違いますが、2001 年に始まった人文学のプロジェクト「Digital Silk Road」です。「Digital Humanities（デジタル人文学）」とは、人文学者と情報学者が協力してやるというものです（図 9）。Digital



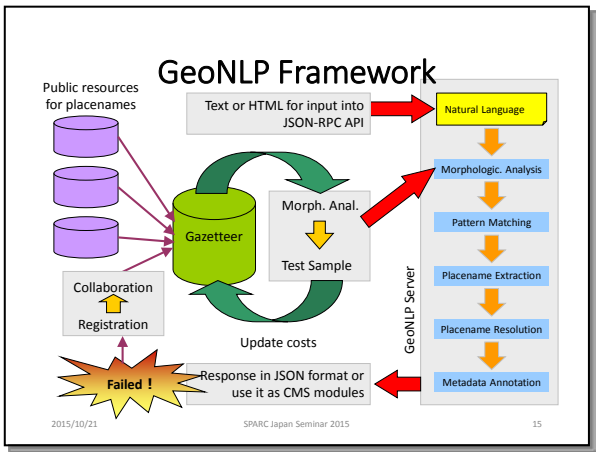
（図 4）



（図 6）



（図 5）

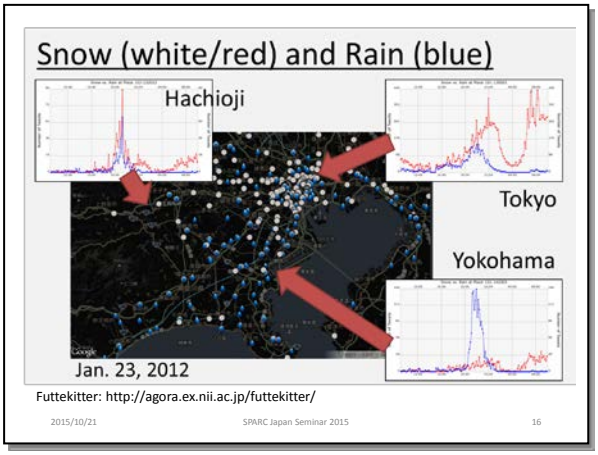


（図 7）

Silk Road もテキスト、マップ、写真、地名データベースなど、いろいろなデータを扱っています。これらをどう使うかについて説明します。

これは一つの例ですが、シルクロードの地図をデジタル化して貼り付けるところまでは簡単にできます（図 10）。貼り付けてみると、何かよく分からない遺跡があることが分かってきます（図 11）。青い点は有名な遺跡で、地図上に「○○遺跡」と書いてあり、それがどこか分かりますが、赤い点は地図上に「○○遺跡」と書いてあっても今はどこにあるか分からないというものです。こういうものが地図をデジタル化すると見つかってきます。なぜ分からないかというと、地図に誤差があって、「○○遺跡」と書いてある場所に、今行ってもないけれど、地図の誤差を考えて行くと遺跡が見つかるという状況にあるからです（図 12）。地図と写真などを組み合わせて現在地を推定する手法をつくり、遺跡を発見するというプロジェクトを行いました。これで実際、遺跡が見つかって、昔の本に書いてあるこの遺跡は、今のこの遺跡だということが分かってきたのです（図 13）。

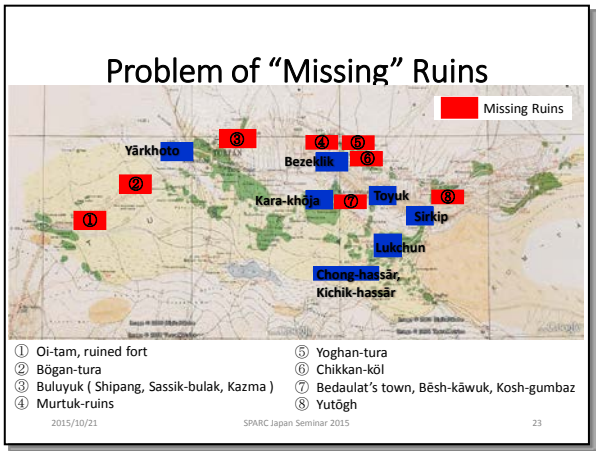
そのような面白い成果が出たのですが、最初は批判されました。特に人文学者から、「何をやっているのか分からない」「今までの仕事とあまりにも違っていて評価できない」など、いろいろな批判を受けました。最初は誤差がどうだといったテクニカルな言葉で説明をしていたので全然分かってもらえなかったのですが、その後に人文学で使われている史料批判という概念を用いて、史料批判をデジタル化して行ったのだと言うと、ふっと分かった人が出てきて、これはいい手法だと言われるようになりました。もちろん理解できない人もいるのですが、きちんと理解できるようにするには言葉遣いを変えることが非常に重要で、コミュニテ



（図 8）

（図 10）

（図 9）



（図 11）

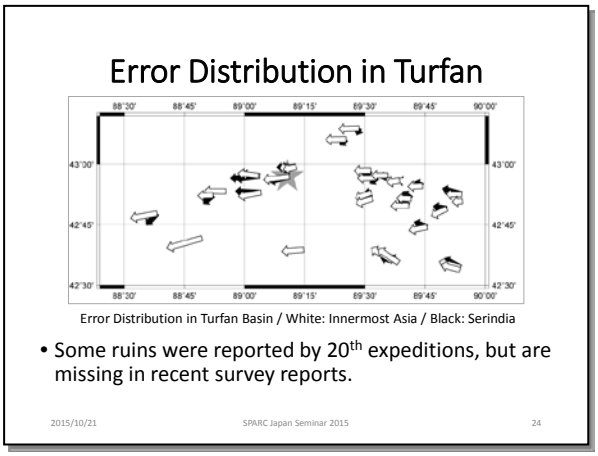
ィー間のバリアをまたぐには、どのように説明するかが重要であるということが分かりました。

このような経験をして、バリアを壊すような研究はどういうものかと考えました。コミュニティーを横断する研究は、データをもらってきて、自分が知っている手法を当てはめて、結果を出すというのが通常のやり方です。これもいいのですが、これでは、既に知られている結果を再現しているだけだったり、正しいけれど当たり前だという結果が出たりすることがあります。そうではない研究にするには、問題について深く理解して、研究のやり方そのものを新しくすることがキーポイントだと思います。そうすると、コラボレーションしないと得られないような結果が得られるのではないでしょうか。そのようなバリアがある中での研究にはいろいろなやり方があると感じています。

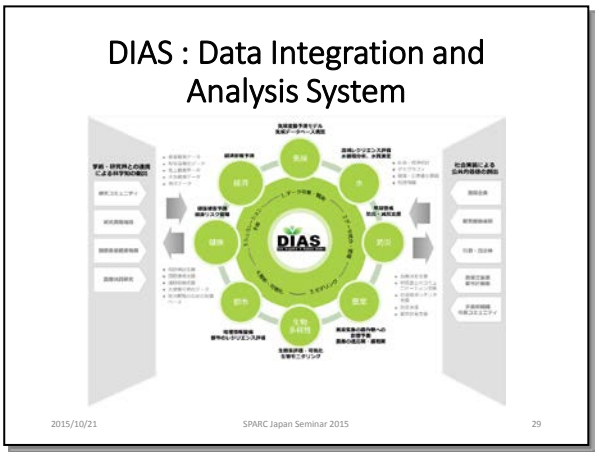
バリアをどう壊すか

バリアを壊すような研究をするときには、やはりデータは重要だと思っています。その一例として DIAS（Data Integration and Analysis System）というプロジェクトを紹介します。これは、地球環境データを 1 カ所に集め、気候、水、防災、農業、生物多様性、都市、健康、経済といったいろいろな分野の研究者がそのデータを共同で分析してあれこれ議論することで新しい成果を得るというプロジェクトです（図 14）。これがいいのは、同じデータについて議論することによって、知見の共有などができていくことです。

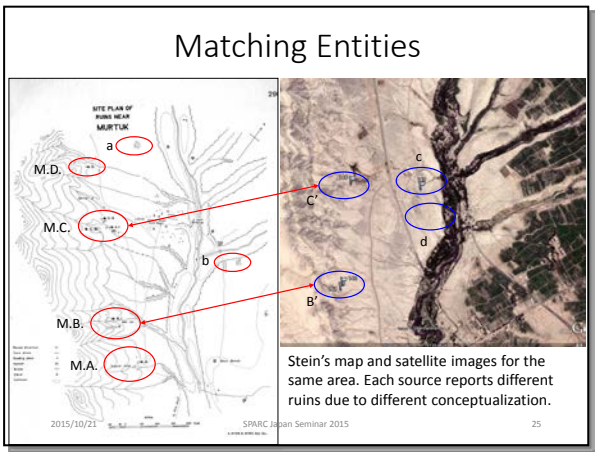
DIAS（図 15）はデータをアーカイブするだけではなく、人が集まってそのデータについて共同で作業するコミュニティーのインフラストラクチャーと、データのインフラストラクチャーがあって、それが合わさって DIAS というインフラストラクチャーができてい



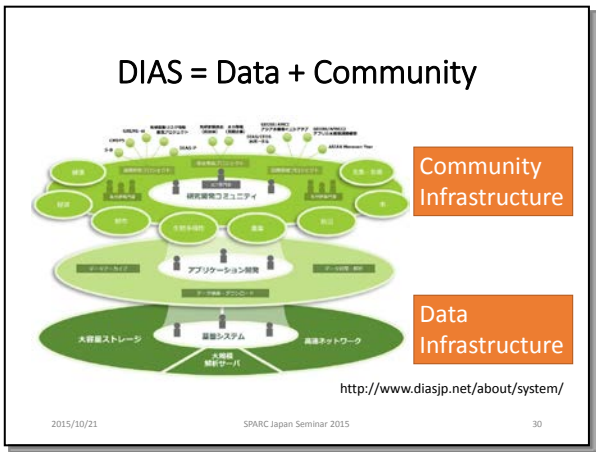
（図 12）



（図 14）



（図 13）



（図 15）

るという 3 層構造になっています。

データをシェアすると、データについてあれこれ分析したりディスカッションしたりということができるので、いろいろなものが集まってきます。一方、例えば知識を共有するとなると、知識は基本的に使うだけなので、そこに集まって何かやるということにはあまりなりません。ですから、データを集めると、そこにコミュニティーが集まってくる。ナレッジだと何か分散していくようなイメージが私にはあります。

重力のようにどんどん集まってくるという動きがつくれれば理想的ですが、どういうものがそのような力を生み出すかを考えました。まず、長くやっているサステイナブルなインフラストラクチャーがあると、ここは信用できるという気分が生まれて、そこに集まる力ができます。もう一つ、単純にデータの量が多くなると、このデータにこのデータを結び付けると面白いという発想がどんどん生まれ、そこに集まってきます。そこに重力でどんどん集まってくると、データとコミュニティーがたくさん集まってくるというポジティブなサイクルをつくるところまでいけると、このようなインフラストラクチャーがどんどん成長するプロセスに入ると思います。

あなたの中にあるバリア

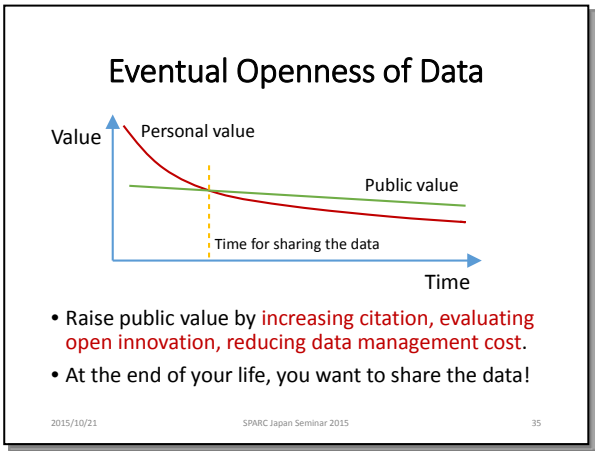
データには、政府データ、研究データ、ビジネスデータがあります。政府データは open by default、原則オープンでなければいけません。ビジネスデータは open by strategy、戦略的にオープンにする、メリットがあればオープンにされます。研究データは多分この中間にあって、eventually open、最終的にはオープンになるけれど、いつオープンにするのかが問題だというエリアにあると思います。

データの価値は、取ったときが一番高く、それから下がっていきます（図 16）。自分にとっての価値（personal value）と社会にとっての価値（public value）があり、恐らく、自分にとっての価値は急激に下がっていき、社会にとっての価値はゆっくり下がっていきます。これについて考えさせられるのは、最近「自分が死ぬ前にデータをオープンにしたい」という相談を受けるときです。もう引退された先生から、「自分が死んだら、これはもう誰も使えなくなるから何とかオープンにしたい。でも、大変過ぎてできない、何とかしてくれ」という話をよく聞きます。

自分にとっての価値は死ぬときは恐らくゼロになってしまいます。そのデータを自分の墓まで持っていきたいわけでないのなら、やはりどこかの時点でオープンにしなければいけないということです。自分にとっての価値と社会にとっての価値が交わるところがオープンにするタイミングなのですが、社会にとっての価値を上げていくと、これがどんどん早まっていきます。例えば、サイテーションをする、イノベーションの価値を言う、データマネジメントコストを下げるなど、いろいろなやり方があると思いますが、とにかくこれをできるだけ早くしていくということです。

今、データは、クローズとオープンのどちらかという二分法になっているように感じますが、リサーチデータについては、このバリアをどんどん下げていくということが重要で、それは Barrier-Less Data という言い方ができると思います。つまり、バリアを今までより下げるならそれは価値があることだという形で、最終状態としてのオープンを言うだけではなく、バリアをどれだけ下げたかという話をするべきだと思います。

バリアを下げるということは、今のシステムに満足していたら別にやらなくていいことなのです。ですか



（図 16）

ら、やはり今に満足していないということが一つ大きな要因としてあると思います。データをシェアして、それについてディスカッションすることは、コラボレーションを促進する上で非常に有効だと思っています。バリアを砕くには、現状に満足せずに、長期的な視点で行動することが重要だと考えています。