Vol.2014-DBS-159 No.19
Vol.2014-IFAT-115 No.19
2014/8/2

情報資源メタデータの評価しやすさの影響要因に関する予備的実験

：書誌データを対象として

矢代寿寛[†1†2]　宮澤彰[†2†3]

書誌などの情報資源メタデータは 1990 年代から Web 公開されているが，現在のオープンデータに求められる二次利用性までは配慮されていなかった．二次利用には目的に応じたデータ評価が必要となる．二次利用を促すため，評価しやすさ/しにくさへの影響要因を明らかにする評価実験を行っている．予備的実験として，延べ 19 名による書誌データ評価を行った結果，被験者属性よりも評価タスクの影響が強いのではないかという仮説が得られた．

Preliminary experiment for analysis of factors that may influence ease of evaluating in information resource metadata: Studies of bibliographic metadata

YASHIRO, Kazunori[†1†2] MIYAZAWA, Akira[†2†3]

Metadata about culturally information resources are Web publishing since the 1990s. But, lacked prospects for the secondary use of the Open Data. To the secondary use, it is necessary to evaluate that value from some facets (quality, structure and others) to serve the purposes of developers. To support the secondary use, conducting the evaluation experiments and collecting data of factors that may influence ease of evaluating. As a result of preliminary experiments by total number of 19 subjects, raise the hypothesis that Impact of evaluation tasks is stronger than subjects attribute.

†1 秋草学園短期大学文化表現学科/Akikusa Gakuen Junior Collage
†2 総合研究大学院大学複合科学研究科情報学専攻/The Graduate University For Advanced Studies(SOKENDAI)
†3 国立情報学研究所情報社会相関研究系/National Institute of Informatics

1. 研究背景

1.1 オープンデータの推進

公的機関により二次利用のために公開・共有されたデータ（とメタデータ），いわゆるオープンデータ（と Linked Open Data）の活用が，わが国でもアプリ開発コンテストやアイデアソンを通じて推進されている．統計や地理空間・施設，資料などのデータは従来，当該分野の専門家が主に利用してきた．しかし，オープンデータ化によって，他分野の専門家やアプリ開発者などの非専門家も以前より目に見える形で利用するようになっている．例えば，経産省らが主催した「オープンデータ・アイデアソン」では，受賞作の開発者にフリーランスや自営業の肩書をみることができる[1]．G8 のオープンデータ憲章の合意と行動計画の策定に伴い，今後も非専門家の利用機会は増すと考えられる[2]．

1.2 オープンデータの二次利用における評価

アプリ開発のような本来の目的ではない二次利用をするためには，自らの目的や開発するアプリの仕様にデータが合い，有効に利用可能か評価する必要がある．この時の評価では，データの特性・品質，形式・構造，利用条件などの事実を特定し（Factual premises），その結果から何らかの基準・指標で価値を判断する（Value premises）[3]．例えば，電力会社により逐次公開される消費電力量を半ばリアルタイムで可視化するアプリの開発において，データが毎時間の消費量グラフの画像ファイルであったときに，「求める時間ごとの消費量がある（特性・品質）」「当事者によるのでおそらく正確である（特性・品質）」「形式

Vol.2014-DBS-159 No.19
Vol.2014-IFAT-115 No.19
2014/8/2

は画像のみでテキストは埋め込まれていない（形式・構造）」「著作権表示しかないが活用は推奨されている（利用条件）」といった事実特定が行われる．そして「形式が合わない．画像からテキストに変換した方が処理しやすいので，変換して利用する」といった価値判断が行われると考えられる．価値判断に加え形成・改善を伴う評価（Formative evaluation）もあるが，この場合には該当しない[4]．

1.3 二次利用における評価の障壁

前述したような評価が実際に滞りなく行われるとは考え難い．公開側（専門家）と二次利用側（非専門家）の間で，情報非対称性（に類する情報量と情報収集機会の差）があると考えられるからである．OECD の Quality Framework など多くのデータ品質基準に含まれる正確性（Accuracy）を評価基準の一つとする場合を例示する[5]．公共施設情報を扱うときに，自治体職員などの現地の関係者であれば，データと実態を比較し正確性の事実を特定することが比較的容易といえる．一方，遠隔地の開発者には実態把握が比較的困難であり，データ解析や凡例との比較，他情報源との対照などから正確性を推察せざるを得ないであろう．情報非対称性という障壁が二次利用側の評価しにくさを生じているといえる．

情報非対称性の影響は，形式・構造の一部であるデータ項目（名称やエレメント，属性）にも及ぶと考えられる．例えば，辞書的に「仏像の大きさ」とされる項目「法量」は，実際には仏像以外の博物館資料の寸法にも利用されている[6][7]．この状況は公開側には適切と評価されるが，辞書的な定義しか知らない二次利用側には誤りに映るであろう．こうした評価の不一致は，特に凡例や API 仕様が不十分でエレメントの定義が不明確であるときに起こりやすいと考えられる．また，非対称でない同分野の専門家が二次利用側であっても，定義の解釈が分かれる曖昧なエレメント（例えば，メタデータスキーマ Dublin core の Coverage）の存在で，評価の不一致は生じうる．公開側に解釈を合わせなければ適切な評価が行えないのであれば，そのコストは障壁であり，評価しにくさを生じるといえる．

加えて，データクレンジングや評価基準・指標の設定なども，必要となれば障壁となり評価をしにくくするであろう．

データの形式・構造面の障壁に関しては，データ公開側に対するガイドラインの整備とデータ評価指標（シグナリングや第三者認証）の開発，二次利用側に対する評価フレームワーク・評価モデルの開発と自動化などにより緩和することが期待される[2][8][9][10]．情報非対称性についても，二次利用側の情報量を増やす効果がある標準化や第三者認証により，ある程度解消を見込むことができる．

1.4 情報資源メタデータの評価に伴う障壁

図書館や博物館の情報資源メタデータは，1993 年頃から Web を通じて既に公開されていたが，近年は新たなオープンデータと組み合わせる形での二次利用が試みられたりしている[11][12][13]．しかし，既存の情報資源メタデータは，オープンデータに求められる二次利用までは配慮されずに作成・公開されてきた．近年はオープン化が進められつつあるとはいえ，評価のしにくさを生む障壁は未だ他のデータよりも多いと考えられる[14]．情報非対称性の例では，データの情報源である現物の確認が，特に博物館資料で困難なことが挙げられる．他にも，「責任表示」「タイトル関連情報」のような曖昧さを残すエレメントが用いられている点，ライセンスの明記がなく二次利用可能かは読み取れない点，国立機関を除くと API やデータセットの公開もほぼみられず技術的な基盤も乏しい点，博物館では標準化・データベース化自体が遅れている点，などが挙げられる[15]．拙稿だけでなく湯佐らの評価でも類似の指摘がなされている[16]．

1.5 評価のしやすさ/しにくさ

業務の範囲内であるためか，わが国では情報資源メタデータの評価を行った研究が乏し

Vol.2014-DBS-159 No.19
Vol.2014-IFAT-115 No.19
2014/8/2

いため，前述した障壁や評価しにくさは認知されているか不明である．また，他にどのような障壁や効果（代表的なハロー効果や系列効果など）が存在するかも明らかではない．仮に，インフォグラフィック化アプリ開発のために，ガイドラインに則って作成され，評価指標により優れたオープンデータと認められた（Open Data Certificate の Expert や 5 star Linked Open Data のような）データを，標準的なフレームワークを用いて評価する，という状況で二次利用側が評価しにくさを感じた場合，形式的な環境整備だけでは障壁の緩和策として不足しているといえる．何が評価のしやすさ/しにくさに影響を与えているか明らかにされなければ，対策が効果的であるとはいいがたい．

オープンデータ（Linked Open Data）に関する評価の研究は，Zeveri らのサーベイにまとめられているように，2000 年代前半からみられ，2010 年以降増加している[17]．一方で，評価のしやすさ/しにくさに関する研究は，オープンデータに関して直接先行するものがみられない．近い領域として，映像と音声における主観品質評価法を評価するメタ評価の研究が行われている．富永らは各評価手法の評価のしやすさを調査した[18][19]．2009 年の実験では時間が，2010 年では評定尺度数がそれぞれ評価しやすさに影響していると推測されている．この他，既存の評価と新たに開発した手法・ツールの比較において評価しやすさが調査される例もみられるが，新手法・ツールの有効性の主張にとどまる．

2. 研究目的

本研究は，オープンデータの評価しやすさ/しにくさへの影響要因を明らかにし，原因となる障壁を緩和し，二次利用を促進することを目標としている．現在は，既にある程度形式面が整備された状態での評価しやすさ/しにくさに影響する要因を明らかにすることを実験により試みている．

本稿では，実験設計のための予備的実験として行った書誌データの評価実験結果から明らかになったことを述べる．仮に，被験者属性の一つである評価の知識や経験が強い影響要因であると明らかになった場合，対策として教育・訓練や評価者の選別を挙げられるが，結論として有用とはいえない．そこで，予備的実験は，被験者属性以外の要因を評価データとして収集する方法について検討しながら適宜修正を加えて行った．この結果から実験は，予備的実験，質的実験，量的実験の三段階で行うものとした．予備的実験により詳細に実験を設計し，質的実験により仮説を導出し，量的実験により仮説を検証する．

3. 研究方法：予備的実験

3.1 実験概要

実験設計のための予備的実験を二度に分けて行った．第 1 次実験は 2013 年 9 月から 10 月に，第 2 次実験は 2014 年 2 月から 4 月に行った．被験者は第 1 次が短期大学生（学生）14 名（うち 1 名は分析から除外）と大学図書館職員（図書館員）1 名の 15 名であった．ある程度属性を揃えて行った第 2 次は学生 2 名と図書館員 3 名の 5 名であった．被験者のうち，学生と図書館職員 1 名ずつが第 2 次実験に再参加しているため，延べ数で 19 名分の実験データを収集した．

評価前の 5 件法による属性アンケートで被験者の知識と経験の程度を Q1 から Q8 まで 8 問で調査した．Dreyfus らの古典的な 5 段階モデル（Novice--Expert）と対応付けるため，選択肢 3 を中間とする評定尺度ではなく累積尺度に近い水準の表現とした[20]．

被験者の知識に関する属性を，順序尺度で表 1 に示す．知識に関して，3 以上と 2 以下が同数になることを企図して被験者を募集したが，図書館と図書は 3 以上（知っている）が，目録とメタデータは 2 以下（少し知っている--知らない）が大半となった．経験に関する属性のうち，大半が全て水準 1（全くしたことがない）の未経験者であったため，それ以外の被験者のみ抜粋して表 2 に示す．

Vol.2014-DBS-159 No.19
Vol.2014-IFAT-115 No.19
2014/8/2

3.2 実験内容

実験で被験者は，既存の書誌データのレコード（書誌レコード）と，その情報源となった現物資料（図書）の比較評価タスクを行った．第 1 次実験では一人あたり 11 件以上，第 2 次実験では 12 件以上のタスクを 1 時間目安で行った．各タスクには意図的な誤りの混入などの調整を加えた．第 1 次ではタスクのセットから 3 件のみ各被験者共通で評価し，第 2 次ではタスクの影響も調査するため 11 件共通で評価した．

書誌レコードは，事実上の標準といえる国立国会図書館の JAPAN/MARC に基づいており，A4 用紙に印刷したものを提示した．評価は質問紙で行った．質問紙には，表 3 の評価項目について 5 段階の水準（5 件法の 3 を中間とする評定尺度）があり，被験者は各項目の 1 つずつの水準を選択することで評価した．第 2 次実験では，a）一般タスクに加え，b）電子書籍を対象に含めるなど複雑化したタスク，c）複数書誌データのコレクション（書誌レコードのまとまり）に対するタスクも追加して行った．コレクション 2 件に対しては，単独の書誌レコードの評価項目に加え，C7）一貫性（Logical Consistency），C8）可読性（Human Readability），C9）完全性（Completeness），C10）有用性（Availability），の 10 項目で評価を行った．各項目名を「これは正確ですか？（正確性）」のように質問文に含んでいる．

表 1 知識に関する属性

被験者		職業*	図書館知識**	図書知識**	目録知識**	メタデータ知識**
第1次	S1	1	2	3	1	1
	S2	1	3	3	1	1
	S3	1	2	3	2	2
	S4	1	3	3	2	1
	S5	1	2	3	2	1
	S6	1	1	1	1	1
	S7	1	2	2	1	1
	S8	2	3	NA	2	3
	S9	2	3	3	2	2
	S10	3	3	4	3	3
	S11	3	3	3	2	2
	S12	3	3	2	2	2
	S13-1	2	3	3	2	2
	L1-1	5	4	4	4	3
第2次	S13-2***	3	3	4	3	3
	S14	3	4	4	3	3
	L1-2***	5	4	4	4	3
	L2	4	3	3	2	1
	L3	5	4	4	4	3

*1）司書科目未履修，2）履修中 1 年，3）履修中 2 年，4）無司書資格図書館員，5）有司書資格図書館員****
**1）全く知らない，2）少し知っている，3）知っている，4）よく知っている，5）とてもよく知っている
***第 2 次実験に再参加
****ここでは名義尺度ではなく順序尺度とみなしている

表 2 経験に関する属性（抜粋）

被験者		メタデータ評価*	書誌調整*	蔵書評価*	図書館評価*
第1次	S11	2***	1	1	1
	L1-1	1	5	5	1
第2次	S13-2**	2	1	2****	1
	L1-2**	2	5	5	2
	L2	1	1	1	1
	L3	1	5	1	1

*1）全くしたことがない，2）1〜2 回した，3）3 回以上したが今はしていない，4）3 回以上したことがあり，年 1〜2 回している，5）毎年 3 回以上している
**第 2 次実験に再参加
***演習科目で経験
****図書館実習で経験

表 3 使用した評価項目

項目	対応概念	第 1 次	第 2 次
正確性	Accuracy	○	○
詳細性	Precision	○	○
可読性	Understandability	○	×
適時性	Timeliness	○	○
信頼性	Credibility	×	○
適合性	Conformance	×	○
標準化	Validity	×	○

評価項目に加えて，各評価対象の図書についての知識と主題についての知識のアンケートを 5 件法で質問した．対象知識の水準は「1）知らない」から「5）内容を覚えている」までとし，主題知識は被験者属性での知識と同じとし，いずれも累積尺度に近い表現をとった．

さらに，評価のしやすさについても 5 件法の評定尺度で質問した．第 1 次実験では「C11 評価は難しい」（「簡単さ」）という設問（1 難

Vol.2014-DBS-159 No.19
Vol.2014-IFAT-115 No.19
2014/8/2

しい--5 簡単）のみであったが，第 2 次実験では，黙従傾向の影響を考慮し，C11 の尺度を逆転項目にした上で，「C12 評価はよくできた」（「出来」）の設問（1 思う--5 思わない）を追加した．タスクの最後に，全体に対する「Q9 評価は難しい」「Q10 よくできた」という設問も用意した．

なお，設問の表現はしやすさ/しにくさの類似表現を学生 15 名から調査した結果による．

4. 研究結果

4.1 評価結果の概要

作業状況に応じて個別に予備タスクを加えた結果，第 1 次実験では 174 件（1 名 11 件，7 名 12 件，6 名 13 件），第 2 次実験では 52 件（3 名 12 件，2 名 13 件）の評価結果が得られた．うち 10 件がコレクションに対する評価であった．

各被験者の水準ごとの評価結果から，第 1 次実験の正確性の項目を抜粋して図 1 に示す．5 が 39.7%（N=174），4 が 36.2%，3 が 9.8%，2 が 14.4%，1 が 0%選択された．被験者を職業 1 の群と 2 以上の群でまとめた上で傾向の違いを分析した結果，χ^2＝8.87 と 5%水準でも有意差はなかった．他の項目も似た傾向であり，第 1 次全体（N=624）として，5 が 31.9%，4 が 36.9%，3 が 10.7%，2 が 16.0%，1 が 4.5%選択された．他の被験者属性での分類では，半数ずつに近いメタデータ知識が 1 か 2 以上かの群同士で，正確性評価などに 1%水準の有意差があった．

評価結果の分析は本旨から外れるため，第 2 次のみと第 1 次・2 次全体については省く．

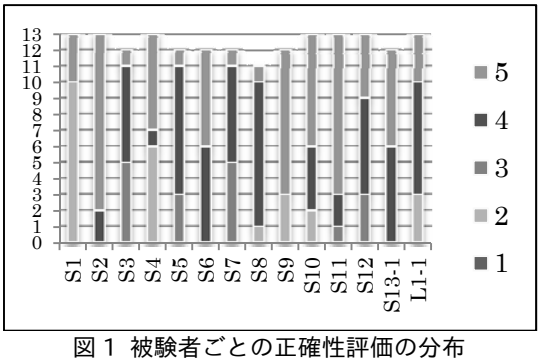

図 1 被験者ごとの正確性評価の分布

本旨である各被験者のタスクごとの評価しやすさ/しにくさの分布を，第 1 次は図 2 に，第 2 次は図 3 に示す．5 の水準が「簡単」または「出来が良い」という評価であり，値が大きいほど「しやすい」ものとみなす．

第 1 次では，5 が 22.5%（N=173，記入ミス 1 件除く），4 が 39.3%，3 が 22.0%，2 が 15.0%，1 が 1.2%選択された．第 2 次では，「C11 簡単さ」の 5 が 22.6%（N=62），4 が 27.4%，3 が 32.3%，2 が 16.1%，1 が 1.6%選択された．単純に比較すると，第 2 次では 4 が減少し，3 のどちらともいえないが増加した．両実験に参加した被験者 S13（学生）と L1（図書館員）は，1 次･2 次間で 5%水準の有意差があった．

第 2 次の「C12 出来」は，5 が 4.9%（N=61，記入ミス 1 件除く），4 が 9.8%，3 が 59.0%，2 が 19.7%，1 が 6.6%選択された．学生と図書館員の群では，1%水準で有意差があった．

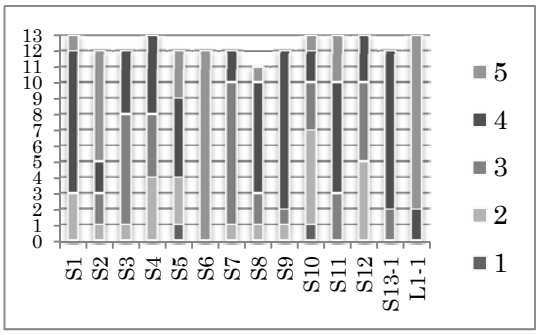

図 2 第 1 次の評価しやすさ/しにくさ

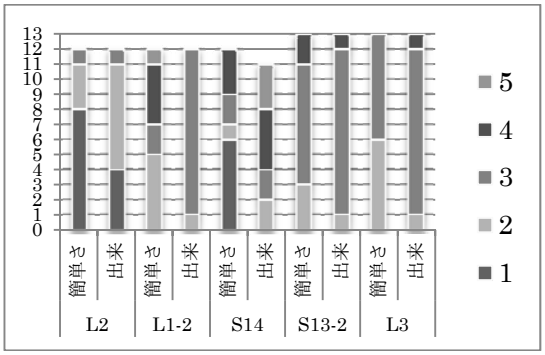

図 3 第 2 次の評価しやすさ/しにくさ

4.2 評価しやすさ/しにくさとの関連

評価しやすさ/しにくさの値について，属性や他の要因と関連があるか分析した．

各知識と経験の値と全体の「Q9 簡単さ」と

Vol.2014-DBS-159 No.19
Vol.2014-IFAT-115 No.19
2014/8/2

「Q10 出来」のポリコリック相関係数（R3.1，polycor パッケージの polychor 関数，2-Step）は，いずれも-0.1 前後とほぼ相関はみられなかった．

他の要因での関連としては，全体の「Q9 簡単さ」と作業時間（の中央値）がポリシリアル相関係数（R3.1, polyserial 関数）0.579 と正の相関がみられた．加えて第 2 次実験では，タスクの種類を a）一般，対象が電子書籍であったり現物でなく書影のみであったり等の b）特殊，c）コレクション，に分類したとき，クラメールの独立係数が「Q9 簡単さ」で 0.38，「Q10 出来」で 0.26 と弱い相関を示した．Q9 と Q10 の合計得点を総合的な「しやすさ」尺度（順序尺度）と仮定した場合にはより強い相関を示す．しかし，a）一般から c）コレクションへのタスクの変化に沿って，傾向が高得点（「しやすい」）から低得点（「しにくい」）へと移るため，系列効果の影響が疑われる．

5. 考察と結論

評価結果に関しては，規則に準拠して作成された標準的な書誌レコードを評価対象としたことが全体，特に属性の値が高い被験者群の高評価に影響したと考えられる．他の検定結果より，概ね，評価結果は不適切ではないと考えられる．

評価しやすさ/しにくさに関しては，まず「Q9 簡単さ」と作業時間（の中央値）に相関がみられたが，「時間がかかるほど評価しやすい」という仮説は不自然と考えられ，別の要因を検討するべきといえる．次に，第 2 次の学生と図書館員の「C12 出来」評価傾向の有意差は，どちらも 1 名ずつが他 3 名と大きく異なる傾向（学生が高評価，図書館員が低評価）を示したためと考えられる．

仮に，経験（と知識）が単純に評価をしやすくするのであれば，第 1 次・2 次両実験に参加した被験者 2 名のしやすさに関する値は増加するはずである．しかし，2 名とも値は減少した．属性のみに要因を限定するのであれば，経験が単純には影響しない，または Einstellung のような抑制的効果を生んでいる可能性がある．属性以外に要因を求めるのであれば，タスクの種類の変化と評価尺度の増加が考えられるものの，現状では不明瞭である．第 2 次ではさらにタスクの種類による弱い相関がみられた．第 2 次はある程度知識に関する属性が共通しているので，他の属性の被験者でも相関がみられるかなどを調査する必要がある．

以上より，現時点での結論（今後の実験における仮説）として，調査した属性よりもタスクの種類による影響がやや強いと考えられる．調査しなかった属性の影響も十分にありうるため，今後は予備的実験の結果を踏まえた質的実験の設計と実施を行う．

参考文献等

[1] 経済産業省，総務省. http://opendata-usecase.go.jp/
[2] 首相官邸. http://www.kantei.go.jp/jp/singi/it2/densi/
[3] Scriven, Michael. The Nature of Evaluation. Part I: Relation to Psychology. Practical Assessment, Research & Evaluation. 6(11). 1999.
[4] 佐々木亮『評価論理』多賀出版. 2010. 167p.
[5] OECD. http://www.oecd.org/std/qualityframeworkforoecdstatisticalactivities.htm
[6] Kotobank（デジタル大辞泉・大辞林第三版）. http://kotobank.jp/word/%E6%B3%95%E9%87%8F
[7] 例えば，九州国立博物館. http://www.kyuhaku.jp/collection/collection_kokuho.html
[8] Open Data Institute https://certificates.theodi.org/
[9] Tim, Berners-Lee. http://www.w3.org/DesignIssues/LinkedData.html
[10] 例えば，ISO/IEC 25012:2008（JIS X 25012:2013）や Bruce, Thomas. Metadata Quality in a Linked Data Context. 2014. http://blog.law.cornell.edu/voxpop/2013/01/24/metadata-quality-in-a-linked-data-context/ など
[11] 農林水産研究情報総合センター. http://ss.cc.affrc.go.jp/ric/opac/opactimeline.html
[12] JST. http://www.jst.go.jp/pr/announce/20131008/
[13] LODAC. http://lod.ac/
[14] 大向一輝. オープンデータと図書館. カレントアウェアネス. 2014. http://current.ndl.go.jp/ca1825
[15] 矢代寿寛，宮澤彰. Web 上で公開された博物館資料メタデータの評価の試み. 情報処理学会研究報告，人文科学とコンピュータ研究会. 2011.
[16] 湯佐安紀子ほか. 文化遺産データベースのメタデータの評価. 日本文化財科学会第 29 回大会. 2012.
[17] Amrapali Zaveri, .et al. Quality Assessment Methodologies for Linked Data: A Survey. Semantic Web journal（Under Review）. 2014.
[18] 富永聡子ほか. モバイル通信用映像に対する主観品質評価法の比較検討. 電子情報通信学会技術研究報告，コミュニケーションクオリティ. 2009.
[19] 富永聡子ほか. モバイル通信における映像品質の主観評価法の性能比較. 電子情報通信学会技術研究報告，コミュニケーションクオリティ. 2010.
[20] Dreyfus, Stuart E., et al.. A five-stage model of the mental activities involved in directed skill acquisition. No. ORC-80-2. 1980.