ISSN 0125-9008 (Print); ISSN 2301-8593 (Online)
DOI: https://doi.org/10.14203/j.baca.v41i1.544
SK Dirjen Risbang -Kemristekdikti No 21/E/KPT/2018 (Peringkat 2 SINTA)

METODE PENILAIAN KUALITAS DATA SEBAGAI REKOMENDASI SISTEM REPOSITORI ILMIAH NASIONAL

Slamet Riyanto[1]*, Ekawati Marlina[2], Hendro Subagyo[3], Hermin Triasih[4], Aris Yaman[5]
[1,2,3,4,5]Lembaga Ilmu Pengetahuan Indonesia

*Korespondensi: slamet.riyanto@lipi.go.id

Diajukan: 13-06-2019; **Direview:** 14-08-2019; **Diterima:** 04-02-2020; **Direvisi:** 08-03-2020

ABSTRACT
High quality data and data quality assessment which efficiently needed to data standardization in the research data repository. Three attributes most used i.e: completeness, accuracy, and timeliness are dimensions to data quality assessment. The purposes of the research are to increase knowledge and discuss in depth of research done. To support the research, we are using traditional review method on the Scopus database to identify relevant research. The literature review is limited for the type of documents i.e: articles, books, proceedings, and reviews. The result of document searching is filtered using some keywords i.e: data quality, data quality assessment, data quality dimensions, quality assessment, data accuracy, dan data completeness. The document that found be analyzed based on relevant research. Then, these documents compare to find out different of concept and method which used in the data quality metric. The result of analysis could be used as a recommendation to implement in the data quality assessment in the National Scientific Repository.

ABSTRAK
Data berkualitas tinggi dan penilaian kualitas data yang efektif dibutuhkan untuk standaridasi data dalam repositori data penelitian. Tiga atribut yang paling banyak digunakan, yaitu kelengkapan, akurasi, dan ketepatan waktu. Tiga atribut tersebut merupakan beberapa dimensi untuk penilaian kualitas data. Penelitian ini bertujuan untuk meningkatkan pengetahuan dan membahas secara mendalam terhadap penelitian yang akan dilakukan. Untuk menunjang penelitian, kami menggunakan metode tinjauan pustaka secara tradisional pada database *Scopus* dan beberapa website terkemuka untuk mengidentifikasi penelitian yang relevan. Studi pustaka dibatasi pada jenis dokumen, yaitu artikel, buku, prosiding, dan tinjauan. Hasil pencarian dokumen disaring menggunakan beberapa kata kunci, yaitu *data quality*, *data quality assessment*, *data quality dimensions*, *quality assessment*, *data accuracy*, dan *data completeness*. Dokumen yang telah diperoleh selanjutnya dianalisis berdasarkan penelitian yang relevan. Selanjutnya, data dianalisis dan dibandingkan untuk mengetahui perbedaan konsep dan metode yang digunakan dalam mengukur kualitas data. Hasil analisis digunakan sebagai rekomendasi untuk diterapkan dalam menilai kualitas data pada sistem Repositori Ilmiah Nasional.

Keywords: Repository; Data; Quality; Data assessment; Research data management; Publication; Indonesia

1. PENDAHULUAN

Menurut perkiraan IDC's "Digital Universe", 40 ZB data akan dihasilkan pada tahun 2020. Munculnya era *big data* menarik perhatian industri, akademisi, dan pemerintah. Para peneliti dan pembuat keputusan secara perlahan menyadari bahwa sejumlah besar informasi memiliki manfaat untuk memahami kebutuhan pelanggan, meningkatkan kualitas layanan, memprediksi serta mencegah risiko. Penggunaan dan analisis *big data* harus berdasarkan pada data yang akurat dan berkualitas tinggi, untuk menghasilkan nilai *big data* (Cai & Zhu, 2015). Peneliti bergantung pada dataset digital dan terkadang mereka menggunakan data yang tidak

dikumpulkan sendiri, tetapi data diperoleh dari sumber publik untuk digunakan kembali (Federer, 2016).

Peneliti membutuhkan infrastruktur yang memastikan aksesibilitas secara maksimum, stabilitas, dan keandalan untuk memfasilitasi bekerja dan berbagi data penelitian (Pampel, et al., 2013). Data merupakan bahan utama pengambilan keputusan operasional, yang bersifat taktis dan strategis. Data merupakan sumber daya penting di semua aplikasi dalam organisasi, bisnis, dan lembaga pemerintah. Kualitas data sangat penting bagi para manajer dan pengambil keputusan untuk menyelesaikan masalah yang terkait dengan kinerja (Batini, Cappiello, Francalanci, & Maurino, 2009). Sebagai produsen, peneliti percaya bahwa mereka ingin memproduksi data yang berkualitas tinggi; dan sebagai konsumen, peneliti ingin memperoleh data yang berkualitas tinggi (Ashley, 2013).

Pada tahun 2016, Pusat Dokumentasi dan Informasi Ilmiah – Lembaga Ilmu Pengetahuan Indonesia (PDII – LIPI) mengembangkan sistem repositori *big data* melalui kegiatan penelitian unggulan yang didanai oleh LIPI, yang disebut dengan Repositori Ilmiah Nasional (RIN). Pengembangan repositori ini bertujuan untuk menyediakan sarana penyimpanan data penelitian yang aman dan handal. Repositori ini juga berfungsi sebagai diseminator karya ilmiah dalam bentuk artikel, prosiding, buku, laporan, dan karya ilmiah yang lain.

Sebagai sarana penyimpanan data penelitian berskala nasional, sistem RIN memungkinkan peneliti dapat mengelola, berbagi, dan melestarikan data penelitiannya. Agar kegiatan berbagi data menjadi efektif, data harus dapat diandalkan, digunakan, mudah ditemukan, diakses, dan disimpan secara permanen untuk jangka panjang (Austin, et al., 2015). Dalam publikasi data penelitian, RIN mengacu siklus hidup *research data repository* dan mendukung proses kurasi data. Publikasi data harus dikaji ulang secara ketat seperti artikel jurnal dalam literatur akademik dan ilmiah, dan data harus dibagi secara terbuka dalam penyimpanan data yang dikurasi. Data adalah fondasi dari semua hal lain yang mengikuti, dan peneliti harus menerima kredit untuk menghasilkan data yang dapat diandalkan (Austin et al., 2015).

Data penelitian didefinisikan sebagai data digital yang menjadi bagian atau hasil dari proses penelitian. Proses ini mencakup semua tahap penelitian, mulai dari pembentukan data penelitian, studi empiris dalam ilmu sosial atau pengamatan fenomena budaya, hingga publikasi hasil penelitian. Data penelitian digital terbagi dalam tipe data yang berbeda, tingkat agregasi dan format data, yang diinformasikan oleh disiplin ilmu penelitian dan metode penelitian yang digunakan. Repositori data dapat mengurangi upaya yang dirasakan, meningkatkan sikap positif, dan mempengaruhi perilaku aktual ilmuwan sosial mengenai penggunaan kembali data. Repositori data memainkan peran penting dalam berbagi dan penggunaan kembali data, meskipun penggunaan kembali data masih terjadi melalui pertukaran dan interaksi antar-orang (Yoon & Kim, 2017).

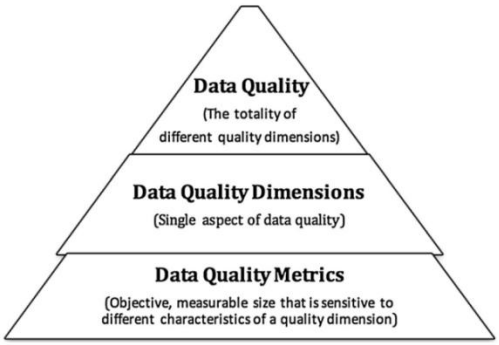
Pada tahun 2009, Komisi Eropa (*European Commission*) menyatakan bahwa lanskap repositori data di seluruh Eropa cukup heterogen. Tetapi, ada dasar yang kuat untuk mengembangkan strategi yang koheren untuk mengatasi fragmentasi dan memungkinkan komunitas riset untuk mengelola, menggunakan, berbagi, dan mempertahankan data dengan lebih baik (Kindling et al., 2017). Pada tahun 2012, proyek pengembangan *Research Data Repository* (RDR) telah menganalisis dan merumuskan kosakata yang mencakup aspek penting dalam repositori data penelitian, yaitu *general information, responsibilities, policies, legal aspects, technical standards*, dan *quality standards* (Spier, et al., 2012). *The Registry or*

Research Data Repository (re3data.org) mendeskripsikan beberapa aspek dalam RDR sebagaimana dijelaskan pada Gambar 1.

Gambar 1. Aspek RDR

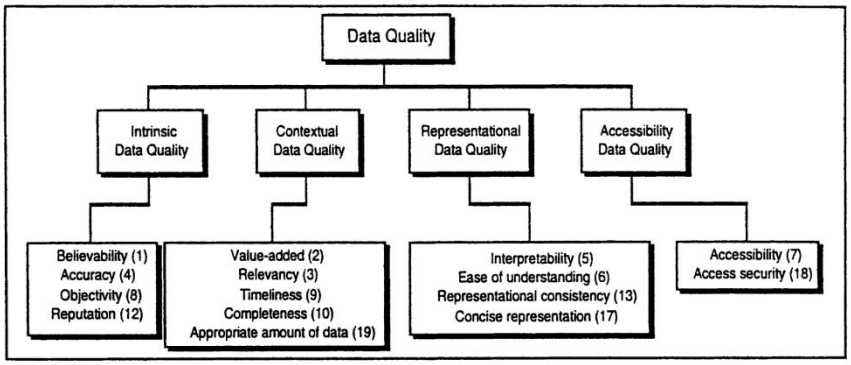
Salah satu aspek repositori yang dibahas dalam tulisan ini adalah *quality standards*, khususnya tentang penilaian kualitas data (*data quality assessment*). Standar kualitas data merupakan standar data organisasi untuk menghasilkan data yang dapat dioperasikan (Zhu & Wu, 2014).

Ada beberapa model siklus hidup data penelitian, diantaranya *Research Data Lifecycle Model, DCC Curation Lifecycle Model,* dan *DDI 3.0 Conceptual Model*. Setiap model memiliki pandangan bahwa satu proses berkaitan dengan kualitas suatu *dataset* (Ashley, 2013). Untuk memastikan standar kualitas data, perlu ada konsensus secara objektif untuk menilai keandalan data (Azeroual, Saake, & Wastl, 2018). Kualitas data dapat dilihat melalui dimensi kualitas data sebagaimana yang tergambar pada piramida berikut ini (Gambar 2).



Gambar 2. Piramida kualitas data (Carey & Ceri, 2006)

Penilaian kualitas data telah dibahas secara intensif dalam penelitian dan praktik. Untuk mendukung manajemen mutu data dan pengambilan keputusan yang berorientasi ekonomi di bawah ketidakpastian, penting untuk menilai tingkat kualitas data menggunakan metrik yang telah ditemukan (Heinrich, Hristova, Klier, Schiller, & Szubartowicz, 2018). Wang & Strong (1996b) telah melakukan penelitian menggunakan metode survei untuk merumuskan dimensi kualitas data (Gambar 3). Kerangka kerja tersebut dapat dijadikan sebagai dasar untuk mengukur, menganalisis, dan meningkatkan kualitas data.



Gambar 3. Konsep kerangka kerja kualitas data

Dimensi kualitas data seperti akurasi, keandalan, ketepatan waktu, kelengkapan, dan konsistensi dapat diklasifikasikan ke tampilan internal dan eksternal. Setiap klasifikasi ini dapat dibagi menjadi dimensi yang berhubungan dengan data dan sistem. Sebagai alternatif, dimensi kualitas data juga dapat diklasifikasikan ke dalam empat kategori, yaitu intrinsik, kontekstual, representasional, dan aksesibilitas (Wang & Strong, 1996a). Tujuan dari penelitian ini untuk mengulas dimensi kualitas data melalui perbandingan dan memilih dimensi yang sesuai dengan perilaku peneliti dalam mengelola data penelitian melalui RIN.

2. TINJAUAN PUSTAKA

2.1 Kualitas Data

Literatur menyediakan berbagai teknik untuk menilai dan meningkatkan kualitas data, seperti hubungan catatan, aturan bisnis, dan ukuran kesamaan (Batini, et al., 2009). Data digambarkan sebagai objek dunia nyata yang dapat disimpan, diambil, dan diuraikan selama proses perangkat lunak dan dikomunikasikan melalui jaringan (Carey & Ceri, 2006). Kualitas data memainkan peran penting dalam kegunaan dan interpretasi data spesifik institusi. Data penelitian dapat diakses secara terbuka melalui tiga strategi, yaitu: (1) publikasi data penelitian, sebagai objek informasi independen, melalui repositori; (2) publikasi data penelitian dengan dokumentasi tekstual yang disebut sebagai makalah data *(data paper*); (3) publikasi data penelitian sebagai pengayaan teks interpretative (*enriched publication*) (Pampel, et al., 2013).

Kualitas data juga merupakan pertimbangan yang signifikan untuk sumber data eksternal (Azeroual, et al., 2018). Kualitas data merupakan serangkaian tindakan yang menentukan apakah data dapat dipahami secara independen untuk dapat digunakan kembali. Penggunaan kembali data berarti bahwa para peneliti asli atau peneliti lain dapat menggunakan data pada waktu mendatang tanpa menentukan apa yang mungkin digunakan secara spesifik (Peer, Green, & Stephenson, 2014). Persyaratan dan kategori kualitas data ditunjukkan pada Tabel 1.

Tabel 1. Persyaratan dan Kategori Kualitas Data

Persyaratan	Kategori
Accuracy	Sejauh mana nilai data sesuai dengan nilai aktual atau nilai sebenarnya
Relevancy	Sejauh mana data berlaku (terkait) dengan tugas pengguna data
Representation	Sejauh mana data disajikan dengan cara yang jelas dan jelas
Accessibility	Sejauh mana data tersedia

Sumber: Wang & Strong (1996b)

Semua aktor yang terlibat dalam setiap tahap siklus hidup data penelitian memiliki kepedulian tentang kualitas data. Namun, para pelaku penelitian jarang menyetujui kualitas data yang dimaksud, misalnya saya menginginkan data yang komprehensif; anda menginginkan data yang tepat waktu; dia menginginkan data yang akurat; dan mereka menginginkan data yang gratis (Ashley, 2013).

2.2 Dimensi Kualitas Data

Dimensi kualitas data merupakan kumpulan atribut kualitas data yang mewakili satu aspek kualitas data (Wang & Strong, 1996b). Karakteristik ontologi dimensi kualitas data sangat penting untuk memberikan pemahaman yang lebih baik tentang kualitas data. Hasil penelitian terdahulu menyarankan bahwa dimensi kualitas data dapat dijelaskan melalui konsep multidimensi seperti akurasi data, konsistensi data, kelengkapan data, dan ketepatan waktu data (Izham, Sidi, Ishak, Suriani, & Jabar, 2017). Dimensi kualitas data dapat dilihat pada Tabel 2.

Tabel 2. Dimensi Kualitas Data

Dimensi	Keterangan
Completeness	Sejauh mana data cukup luas, mendalam, dan ruang lingkup untuk tugas yang dihadapi
Correctness/free of error	Sejauh mana data benar dan dapat diandalkan
Representation	Sejauh mana usia data tepat untuk tugas yang dihadapi
Consistency	Sejauh mana data selalu disajikan dalam format yang sama dan kompatibel dengan data sebelumnya

Sumber: Wang & Strong (1996b)

Metodologi untuk mengukur dimensi dan peningkatan kualitas data seperti *Total Information Quality Management* (TIQM) dan *Total Data Quality Management* (TDQM). Metodologi tersebut fokus pada pencarian penyebab tidak adanya kualitas dari data. Prosesnya dengan mengukur dan meningkatkan kegiatan pada satu set dimensi kualitas (misalnya, akurasi, ketepatan waktu, kelengkapan) dan menganalisis ketergantungan antara dimensi dengan menggunakan sifat entropi (Shariat, et al., 2013).

2.3 Metrik Kualitas Data

Metrik kualitas data diperlukan untuk mengevaluasi dimensi kualitas data. Mereka membentuk ukuran kualitas yang memungkinkan pernyataan kuantitatif. Metrik kualitas ini membentuk basis operasional untuk menentukan kualitas data (Gebauer & Windheuser, 2017). Metrik dipahami untuk menjelaskan metode dan sistem, sebagai sebuah hasil, menyediakan nilai yang dihitung dan angka-angka kunci (Azeroual, et al., 2018). Pengukuran kualitas data tanpa mendefinisikan dimensi yang relevan dengan jelas sebelumnya, tidak mungkin untuk dilakukan (Martin, 2005).

Kesalahan data seperti nilai-nilai yang hilang, duplikat, kesalahan ejaan, pemformatan yang salah dan inkonsistensi (dapat terjadi selama pengumpulan), transmisi dan integrasi informasi penelitian dalam sistem yang berbeda, perlu dikenali lebih awal agar dalam pengelolaannya dapat terlaksana secara efisien (Azeroual, et al., 2018). Sedangkan keberhasilan implementasi kualitas data ditentukan oleh tujuh faktor yaitu: tanggung jawab manajemen, biaya operasi dan jaminan, penelitian dan pengembangan, produksi, distribusi,

manajemen personalia, dan fungsi hukum (Wang, Storey, & Firth, 1995). Penelitian tentang kualitas data dapat dibagi menjadi empat kelompok, yang mencakup: kualitas data untuk situs web, kualitas data untuk mendukung keputusan, penilaian kualitas data, dan aplikasi kualitas data lainnya seperti proses pengembangan perangkat lunak dan kualitas data medis (Xiao, Lu, Liu, & Zhou, 2014). Penilaian kualitas data terkait sistem RIN, mengadopsi kategori yang diusulkan oleh Xiao, et al. Data penelitian yang tersimpan dalam RIN mencakup multidisiplin ilmu, sehingga metrik penilaian tidak spesifik pada bidang tertentu.

3. METODE

Kami menggunakan *literature review* dalam meningkatkan pemahaman untuk merumuskan dimensi kualitas data yang diterapkan pada sistem RIN. Tulisan ini terkait erat dengan literatur yang membahas kualitas data, dimensi kualitas data, dan metrik kualitas data. Untuk menggali informasi secara mendalam, peneliti melakukan pencarian literatur yang relevan agar dapat menjadi rekomendasi penentuan standar kualitas data RDR, khususnya RIN. Pada penelitian ini, peneliti mempertimbangkan tiga dimensi kualitas data, yaitu akurasi, kelengkapan, dan ketepatan waktu.

4. HASIL DAN PEMBAHASAN

4.1 Metode Penilaian Kualitas Data

Metode yang diterapkan dalam RIN mengadopsi dua jenis metode, yaitu *data-driven* dan *process-driven*. *Data-driven* untuk meningkatkan kualitas data dengan cara memodifikasi nilai data secara langsung. Sedangkan, *process-driven* untuk meningkatkan kualitas data dengan mendesain ulang proses yang membuat atau memodifikasi data (Batini, et al., 2009). Berikut ini ada beberapa teknik dalam strategi *data-driven,* yaitu:

1) Akuisisi data baru: meningkatkan kualitas data dengan memperoleh data berkualitas lebih tinggi untuk menggantikan nilai-nilai yang meningkatkan masalah kualitas;
2) Standaridasi atau normalisasi: menggantikan atau melengkapi nilai data yang tidak standar dengan nilai yang standar;
3) Merekam hubungan: mengidentifikasi representasi data dalam dua atau beberapa tabel yang merujuk ke objek dunia nyata yang sama;
4) Integrasi skema dan data: memiliki tujuan yang memungkinkan pengguna dapat mengakses data yang heterogen melalui penyimpanan data terpadu;
5) Sumber dapat dipercaya: memilih sumber data berdasarkan kualitas datanya;
6) Lokalisasi kesalahan dan koreksi: mengidentifikasi dan menghilangkan kesalahan kualitas data dengan mendeteksi rekaman yang tidak memenuhi seperangkat aturan kualitas yang diberikan;
7) Optimalisasi biaya: mendefinisikan tindakan peningkatan kualitas sepanjang serangkaian dimensi dengan meminimalkan biaya (Batini, et al., 2009).

Strategi diatas dapat diadopsi oleh RIN untuk peningkatan kualitas data, karena koleksi yang tersimpan di RIN sebagian besar dalam bentuk publikasi ilmiah, seperti jurnal, prosiding, dan buku, sedangkan untuk data penelitian masih belum banyak yang mendepositkan. Untuk sementara waktu, kendali untuk mempublikasikan data berada di tangan para peneliti sebagai pemilik data. Apalagi RIN mendukung proses kurasi data yang bertujuan untuk menjamin kualitas data ketika dipublikasikan. Selain mengadopsi metode *data-driven*, RIN dapat mengadopsi *process-driven* untuk meningkatkan kualitas data. Karakteristik metode *process-*

driven, yaitu: (1) proses kontrol menyisipkan prosedur pemeriksaan dan kontrol dalam proses produksi data ketika data baru dibuat, dataset diperbarui, atau dataset baru diakses oleh proses tersebut; (2) proses desain ulang untuk menghilangkan penyebab kualitas data yang buruk dan memperkenalkan kegiatan baru yang menghasilkan data dengan kualitas yang lebih tinggi.

Metode lain yang digunakan dalam menentukan penilaian kualitas data adalah melakukan analisis terhadap atribut data publikasi. Ada sekitar 49 atribut yang digunakan untuk menjelaskan kualitas data (Tabel 3). Atribut kelengkapan, akurasi, dan ketepatan waktu merupakan atribut yang paling sering diukur.

Tabel 3. Atribut Kualitas Data

Item	***Attribute***
High data quality (38)	*Completeness, accuracy or positional accuracy, timeliness or up-datedness or currency, validity, periodicity, relevance, reliability, precision, integrity, confidentiality or data security, comparability, consistency or internal consistency or external consistency, concordance, granularity, repeatability, readily useableness or usability or utility, objectivity, ease with understanding, importance, reflecting actual sample, meeting data standards, use of standards, accessibility, transparency, representativeness, disaggregation, data collection method or adjustment methods or data management process or data management*
Poor data quality (11)	*Missing data, under-reporting, inconsistencies, data errors or calculation errors or errors in report forms or errors resulted from data entry, invalid data, illegible hand writing, non-standardization of vocabulary, and inappropriate fields*

Sumber: Chen, Hailey, Wang, & Yu (2014)

a. Metode untuk penilaian dimensi akurasi

Data akurat ketika nilai data yang disimpan dalam *database* sesuai dengan nilai dunia nyata (Batini, et al., 2009). Wang & Strong (1996b) mendefinisikan akurasi data sebagai "sejauh mana data benar, dapat diandalkan dan tersertifikasi". Akurasi adalah ukuran kedekatan nilai data, v hingga beberapa nilai lain, v^1, yang dianggap benar (Batini, et al., 2009). Sebagai contoh jika nama seseorang adalah 'Slamet', nilai v^1 = Slamet adalah benar, sedangkan nilai v = Slamet adalah salah. Ukuran koreksi data membutuhkan sumber referensi otoritas untuk diidentifikasi dan diakses (McGilvray, 2008).

Akurasi data ada dua jenis, yaitu akurasi sintaksis dan akurasi semantik. Akurasi sintaksis adalah kedekatan dari nilai v ke elemen-elemen yang sesuai definisi domain D. Akurasi sintaksis diukur dengan menggunakan fungsi yang disebut fungsi perbandingan, yang mengevaluasi jarak antara v dan nilai dalam D. Edit jarak adalah contoh sederhana dari fungsi perbandingan, dengan mempertimbangkan jumlah minimum penyisipan karakter, penghapusan, dan penggantian untuk mengubah string s ke string s^1 (Batini & Scannapieco, 2016). Sebagai contoh, nama afiliasi "Lemb..ga Ilmu Pengetahuan Indonesia" adalah tidak akurat secara sintaksis karena tidak sesuai dengan nama afiliasi manapun. "Lembaga Ilmu Pengetahuan Indonesia" adalah nama afiliasi yang paling dekat dengan "Lembaga Ilmu Pengetahuan Indonesia", memang jarak edit antara Lemb..ga Ilmu Pengetahuan Indonesia dan Lembaga Ilmu Pengetahuan Indonesia sama dengan 1 dan

hanya sesuai dengan penyisipan karakter "a" dalam *string* Lemb..ga Ilmu Pengetahuan Indonesia.

Keakuratan semantik adalah kedekatan dari nilai v ke nilai v^1 yang sebenarnya. Hal tersebut dapat kita lihat contoh relasi data pada *database* keragaman hayati (Tabel 4). Pertukaran nama *Genus* dalam baris 1 dan 2 adalah contoh dari kesalahan akurasi semantik. Untuk nama *Genus* 1 bernama *Padda* akan diterima, dan itu benar secara sintaksis. Namun demikian, *Padda* bukanlah nama *Genus* untuk *Family Muscicapidae*, sehingga kesalahan akurasi semantik terjadi. Seharusnya nama *Genus* baris 1 dan 2 ditukar posisinya.

Tabel 4. Database keragaman hayati

No	*Family*	*Genus*	*Species*
1	Muscicapidae	*Padda*	*cinerea*
2	Ploceidae	*Pachycephala*	*oryzviora*
3	Nectariniidae	*Leptocoma*	*jugularis*
4	Timaliidae	*Alcippae*	*cinerea*
5	Laniidae	*Pachycephala*	*pectoralis*
6	Corvidae	*Kitta*	*chinensis*
7	Meliphagidae	*Lichmera*	*lombokia*
8	Rallidae	*Poliolimnas*	*cinerea*
9	Acledinidae	*Halcyon*	*cyanoventris*
10	Pycnonotidae	*Chloropsis*	*sonnerati*
11	Columbidae	*Ptilinopus*	*melanospila*
12	Turdidae	*Turdus*	*javanicus*

Contoh di atas dengan jelas menunjukkan perbedaan antara akurasi sintaksis dan semantik. Perhatikan bahwa meskipun masuk akal untuk mengukur akurasi sintaksis menggunakan fungsi jarak, akurasi semantik diukur lebih baik dengan domain <*yes, no*> atau <*correct, not correct*> [4].

b. Metode untuk penilaian dimensi kelengkapan

Data lengkap ketika semua nilai data yang diperlukan tersedia. Selain itu, data harus dapat mewakili nilai nol karena dalam beberapa kasus data mungkin tidak memiliki nilai yang terkait (Bovee, Srivastava, & Mak, 2003). Misalnya, karyawan dengan status tidak menikah atau belum menikah akan mengosongkan kolom nama pasangan. Dalam hal ini, nilai nol pada kolom nama pasangan tidak dapat dianggap sebagai data yang tidak lengkap. Data dapat memiliki nilai nol dan keberadaan nilai nol, tidak seharusnya dianggap sebagai data yang tidak lengkap. Data yang tidak lengkap terjadi ketika nilai nol ditetapkan untuk data yang seharusnya memiliki nilai. Hal ini menunjukkan bahwa proses penilaian kualitas data harus dapat mengidentifikasi penyebab nilai *null* yang ditemukan dalam *dataset* sebelum kelengkapan data dinilai. Nilai *null* memiliki arti umum dari nilai yang hilang, nilai yang ada di dunia nyata tetapi tidak tersedia dalam pengumpulan data. Untuk mengkarakterisasi kelengkapan, kita harus memahami 'mengapa nilai tersebut hilang". Kelengkapan data tergantung pada persepsi dan konteks konsumen data, bukan penafsiran pemilik data (Wang & Strong, 1996b). Kelengkapan data terhadap penilaian kualitas data di RIN, mengacu beberapa studi yang pernah dilakukan sebelumnya. Pemeriksaan

kelengkapan data, mengidentifikasi semua *file* yang ada – data, dokumentasi, dan kode diperlukan jika tersedia (Peer, et al., 2014). Salah satu pendekatan pemeriksaan kelengkapan data adalah menambahkan informasi atau petunjuk penggunaan data dan mendeksripsikan data pada sistem repositori.

c. Metode untuk penilaian dimensi ketepatan waktu

Ketepatan waktu data mengacu pada usia data (Wang & Strong, 1996b). Ketepatan waktu data dapat dilihat sebagai atribut tanggal (Bovee, et al., 2003). Atribut tanggal termasuk usia dan *volatilitas* sebagai ukuran ketepatan waktu data. Ketepatan waktu dan tanggal harus diukur oleh pengguna dalam konteks tujuan aplikasi. Ketepatan waktu data sangat penting karena data terkini memiliki potensi yang lebih besar untuk dipertimbangkan sebagai kualitas data yang tinggi (Wang & Strong, 1996b). Data yang didepositkan di RIN mengacu kebijakan Repositori dan Depositori yang diatur dalam Peraturan Kepala LIPI No.12 Tahun 2016, Pasal 9 menyebutkan bahwa "setiap data primer yang dihasilkan dari penelitian dan/atau pengembangan, survei, atau pemikiran sistematis yang dilakukan oleh LIPI maupun pihak lain yang bekerja sama dengan LIPI wajib disimpan dalam depositori LIPI setelah kegiatan berakhir".

Mengacu peraturan tersebut, setiap peneliti atau satuan kerja wajib menyimpan data penelitiannya ke sistem RIN setelah kegiatan berakhir. Idealnya, kegiatan penelitian yang menghasilkan data tidak harus menunggu sampai kegiatan penelitian berakhir, tetapi setiap saat selama kegiatan penelitian berlangsung. Setiap peneliti diwajibkan melaporkan perkembangan penelitiannya, termasuk data yang telah dihasilkan. Ini berkaitan erat dengan keamanan data, semakin cepat dilakukan *back-up* data maka semakin aman. Jika setiap peneliti mengelola data penelitiannya dengan baik dalam sebuah sistem, maka data tersebut dapat digunakan kembali oleh peneliti tersebut, kelompok penelitian, dan lembaga. Hal ini bertujuan untuk mengklaim dan memberitahukan kepada dunia bahwa dia telah melakukan penelitian dengan data tertentu. Semakin lama mempublikasikan datanya, maka akan mengurangi nilai data tersebut – karena perkembangan teknologi yang semakin canggih dan data semakin kompleks.

4.2 Peluang, Tantangan, dan Kendala

Salah satu peran seorang kurator data adalah memeriksa akurasi data yang tersimpan dalam sistem repositori sehingga memenuhi kriteria data yang berkualitas sebelum dipublikasikan. Kemampuan melakukan penilaian data sangat diperlukan oleh seorang kurator data, sehingga peningkatan kemampuan sumber daya manusia menjadi aspek yang sangat penting. Upaya peningkatan kompetensi pengelola data penelitian, diantaranya: selalu melibatkan minimal satu orang kurator data ketika penelitian; mengikutsertakan sumber daya manusia dalam kegiatan seminar atau pelatihan (dalam atau luar negeri), terkait dengan manajemen data penelitian, kurasi data, penilaian kualitas data, dan topik lain yang relevan. Kurator data yang sering melakukan penilaian kualitas data, akan memiliki peluang sebagai ilmuwan data (*data scientist*). Sebagai ilmuan data, seorang kurator data tidak hanya menilai data, tetapi juga melakukan analisis data untuk kepentingan peramalam, prediksi, klasifikasi, maupun klasterisasi data berbasis riset.

Data yang tersimpan dalam repositori sudah pasti sangat beragam, mencakup berbagai disiplin ilmu. Keberagaman pengetahuan dan keahlian yang dimiliki staf pengelola repositori menjadi sangat penting, hal ini dikarenakan data yang harus dikelola sangat beragam.

Pengetahuan tentang metadata juga perlu mendapat perhatian khusus, tidak semua orang memiliki kemampuan memahami berbagai metadata standar.

Ada kalanya seorang pemilik data menyerahkan datanya ke pengelola repositori secara langsung. Hal tersebut bisa terjadi karena tidak semua pemilik data memiliki waktu untuk mendepositkan datanya ke sistem repositori. Pemilik data tidak ingin direpotkan dengan hal teknis untuk mendepositkan datanya ke sistem repositori. Ketika pemilik data menyerahkan datanya ke pengelola repositori, umumnya tidak disertai dengan metadata secara lengkap. Padahal kelengkapan metadata sangat diperlukan untuk memberikan informasi terkait data yang disimpan. Metadata standar yang umumnya sering digunakan untuk pertukaran data hendaknya berisi *title*, *creator*, *description*, *subject*, *date*, *publisher*, dan *contributor*.

Penilaian kualitas data bergantung pada pemahaman seorang kurator data dalam mendefinisikan sebuah data yang mencakup *accuracy*, *relevancy*, *representation*, dan *accessibility*. Hal ini menjadi kendala dalam menjamin kualitas data yang dipublikasikan melalui sistem repositori. Oleh karena itu, peningkatan kompetensi sumber daya manusia tentang penilaian kualitas data harus menjadi perhatian khsusus. Kegiatan mengkurasi data bukan hanya sekadar memeriksa metadata, melengkapi data pendukung, petunjuk penggunaan data, maupun aspek legalitas. Namun perlu memahami secara substansi, sehingga data yang dikelola dan dipublikasikan dapat digunakan kembali (*reuse*) dan menghasilkan data baru (*reproduce*).

5. KESIMPULAN

Indikator penilaian kualitas data harus berpedoman pada beberapa dimensi dan tipe data yang diukur. Seorang penilai kualitas data perlu memahami berbagai indikator dalam sebuah data. Kompentensi sesuai bidang penelitian sangat direkomendasikan ketika melakukan penilaian kualitas data. Dua parameter kualitas data, seperti ketepatan waktu dan aksesibilitas biasanya kurang diperhatikan oleh pengelola data penelitian. Dimensi kualitas data yang dapat diadopsi oleh RIN untuk manajemen data penelitian yaitu *accuracy*, *completeness*, dan *timeliness*.

DAFTAR PUSTAKA

Ashley, K. 2013. Data Quality and Quration. *Data Science Journal*, *12*(10), 65–68.

Austin, C., Brown, S., Fong, N., Humphrey, C., Leahey, A., & Webster, P. 2015. Research Data Repositories: Review of Current Features, Gap Analysis, and Recommendations For Minimum Requirements.

Azeroual, O., Saake, G., & Wastl, J. 2018. Data Measurement in Research Information Systems: Metrics for The Evaluation of Data Quality. *Scientometrics*, *115*(3), 1271–1290. https://doi.org/10.1007/s11192-018-2735-5.

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. 2009. Methodologies for Data Quality Assessment and Improvement. *ACM Computing Surveys*, *41*(3), 1–52. https://doi.org/10.1145/1541880.1541883.

Batini, C., & Scannapieco, M. 2016. *Introduction to Information Quality*. *Data and Information Quality: Dimensions, Principles and Techniques*. https://doi.org/10.1007/978-3-319-24106-7_1.

Bovee, M., Srivastava, R. P., & Mak, B. 2003. A Conceptual Framework and Belief Function Approach to Assessing Overall Information Quality. *Proceedings of the Sixth International Conference on Information Quality*, *18*, 311–328. https://doi.org/10.1002/int.10074.

Cai, L., & Zhu, Y. 2015. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal*, *14*. https://doi.org/10.5334/dsj-2015-002.

Carey, M. J., & Ceri, S. 2006. *Data-Centric Systems and Applications Data, 49*. Germany: Springer Berlin Heidelberg.

Chen, H., Hailey, D., Wang, N., & Yu, P. 2014. A Review of Data Quality Assessment Methods for Public Health Information Systems. *International Journal of Environmental Research and Public Health*, *11*(5), 5170–5207. https://doi.org/10.3390/ijerph110505170

Federer, L. 2016. Research Data Management in The Age of Big Data: Roles and Opportunities for Librarians. *Information Services and Use*, *36*(1–2), 35–43. https://doi.org/10.3233/ISU-160797.

Gebauer, M., & Windheuser. 2017. *Structured Data Analysis, Profiling, and Business Rules*. Germany: Springer Fachmedien Wiesbaden.

Heinrich, B., Hristova, D., Klier, M., Schiller, A., & Szubartowicz, M. 2018. Requirements for Data Quality Metrics. *Journal of Data and Information Quality*, *9*(2), 1–32. https://doi.org/10.1145/3148238.

Izham, J.M., Sidi, F., Ishak, I., Suriani, A.L., & Jabar, A. M. 2017. A Review of Data Quality Research in Achieving High Data Quality Within Organization. *Journal of Theoretical and Applied Information Technology*, *95*(12), 2647–2657.

Kindling, M., Pampel, H., Van de Sandt, S., Rücknagel, J., Vierkant, P., Kloska, G., … Scholze, F. 2017. The Landscape of Research Data Repositories in 2015: A Ee3data analysis. *D-Lib Magazine*, *23*(3–4). https://doi.org/10.1045/march2017-kindling.

Martin, M. 2005. *Measuring and Improving Data Quality. Part II: Measuring data quality*. NAHSS Outlook.

McGilvray, D. 2008. Executing Data Quality Projects. *Executing Data Quality Projects*, 256–277.

https://doi.org/10.1016/B978-012374369-5.50006-4.

Pampel, H., Vierkant, P., Scholze, F., Bertelmann, R., Kindling, M., Klump, J., … Dierolf, U. 2013. Making Research Data Repositories Visible: The re3data.org Registry. *PLoS ONE*, *8*(11), e78080. https://doi.org/10.1371/journal.pone.0078080.

Peer, L., Green, A., & Stephenson, E. 2014. Committing to Data Quality Review. In *The 9th International Digital Curation Conference*.

Shariat, P.P. H., Sidi, F., Affendey, L. S., Jabar, M. A., Ibrahim, H., & Mustapha, A. 2013. A Framework to Construct Data Quality Dimensions Relationships. *Indian Journal of Science and Technology*, *6*(5), 4422–4431. https://doi.org/10.1145/2480362.2480718.

Spier, S., Gundlach, J., Pampel, H., Kindling, M., Kirchhoff, A., Klump, J., … Scholze, F. 2012. *Vocabulary for the Registration and Description of Research Data Repositories. Version 2.0*. https://doi.org/10.2312/re3.002.

Wang, R. Y., Storey, V. C., & Firth, C. P. 1995. A Framework for Analysis of Data Quality Research. *IEEE Transactions on Knowledge and Data Engineering*, *7*(4), 623–640. https://doi.org/10.1109/69.404034.

Wang, R. Y., & Strong, D. M. 1996a. Anchoring Data Quality Dimensions in Ontological Foundations. *Communications of the ACM*, *39*(11), 86–95.

Wang, R. Y., & Strong, D. M. 1996b. Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, *12*(4), 5–33. https://doi.org/10.1080/07421222.1996.11518099.

Xiao, Y., Lu, L. Y. Y., Liu, J. S., & Zhou, Z. 2014. Knowledge Diffusion Path Analysis of Data Quality Literature: A Main Path Analysis. *Journal of Informetrics*, *8*(3), 594–605. https://doi.org/10.1016/j.joi.2014.05.001.

Yoon, A., & Kim, Y. 2017. Social scientists' data reuse behaviors: Exploring the roles of attitudinal beliefs, attitudes, norms, and data repositories. *Library and Information Science Research*, *39*(3), 224–233. https://doi.org/10.1016/j.lisr.2017.07.008.

Zhu, H., & Wu, H. 2014. Assessing The Quality of Large-Scale Data Standards: A Case of XBRL GAAP Taxonomy. *Decision Support Systems*, *59*(1), 351–360. https://doi.org/10.1016/j.dss.2014.01.006.