Information quality, data and philosophy
Phyllis Illari and Luciano Floridi

1 Understanding information quality

In this opening chapter, we review the literature on information quality (IQ). Our major aim is to introduce the issues, and trace some of the history of the debates, with a view to situating the chapters in this volume – whose authors come from different disciplines – to help make them accessible to readers with different backgrounds and expertise. We begin in this section by tracing some influential analyses of IQ in computer science. This is a useful basis for examining some examples of developing work on IQ in section two. We look at some cases for applying IQ in section three, and conclude with some discussion points in section four.

1.1 The MIT group

The issue of IQ came to prominence in computer science, where a research group at MIT launched and defined the field of IQ in the 1990s. The MIT group was committed to the idea that considerably more could be done about IQ problems. Members of the group were enormously influential, and generated a large and thriving community: the 18th annual IQ conference was held in 2013 in Arkansas, USA.

The consistent primary message of the MIT group is that quality information is information that is fit for purpose, going far beyond mere accuracy of information. This message bears repeating, as mere accuracy measures are still sometimes informally described as IQ measures. Since the MIT group conceives of IQ projects initially as data management for business, it presses this message as a recommendation to consider 'information consumers': constantly ask what it is that consumers of information need from it, treating data as a valuable and important product, even if the consumers of that product are internal to the organization.

The idea that IQ is a *multidimensional* concept, with accuracy as only one dimension, is now embedded. Much of the MIT group's early work aimed to identify and categorise the dimensions of IQ. This work falls into two different methodological approaches, also identified by Batini and Scannapieco (2006, p. 38).

The first methodological approach is called 'empirical' by Batini and Scannapieco, and by members of the MIT group. Broadly speaking, it consists in surveying IQ professionals, both academics and practitioners, on what they rate as important IQ dimensions, and how they classify them. In the past, some work also examined published papers on IQ, and surveyed professional practitioners at conferences, to

identify important IQ skills. The empirical approach is based on initial work by Wand and Wang in 1996, making a citation count, actually given in Wang (1998). In line with the focus on information users, data consumers were also interviewed (Batini & Scannapieco, 2006, p. 38.).

The categorisation of Wang (1998) is one of the earliest and still most influential categorisations of IQ dimensions. **Error! Reference source not found.** below is a reconstruction of the table given in the paper (Wang, 1998, p. 60):

IQ Category	IQ Dimensions
Intrinsic IQ	Accuracy, Objectivity, Believability, Reputation
Accessibility IQ	Access, Security
Contextual IQ	Relevancy, Value-Added, Timeliness, Completeness, Amount of data
Representational IQ	Interpretability, Ease of understanding, Concise representation, Consistent representation

Table 1: Wang's categorisation (Source: Wang (1998)),

Another important paper is Lee, Strong, Kahn, and Wang (2002), who give us two comparison tables of classifications of IQ dimensions, one for academics reconstructed in Table 2 (Lee et al., 2002, p. 134), laid out according to the Wang (1998) categories, and one for practitioners (Lee et al., 2002, p. 136).

	Intrinsic IQ	Contextual IQ	Representational IQ	Accessibility IQ
Wang and Strong [39]	Accuracy, believability, reputation, objectivity	Value-added, relevance, completeness, timeliness, appropriate amount	Understandability, interpretability, concise representation, consistent representation	Accessibility, ease of operations, security
Zmud [41]	Accurate, factual	Quantity, reliable/timely	Arrangement, readable, reasonable	
Jarke and Vassiliou [16]	Believability, accuracy, credibility, consistency, completeness	Relevance, usage, timeliness, source currency, data warehouse currency, non-volatility	Interpretability, syntax, version control, semantics, aliases, origin	Accessibility, system availability, transaction availability, privileges
Delone and McLean [11]	Accuracy, precision, reliability, freedom from bias	Importance, relevance, usefulness, informativeness, content, sufficiency, completeness, currency, timeliness	Understandability, readability, clarity, format, appearance, conciseness, uniqueness, comparability	Usableness, quantitativeness, convenience of access[a]
Goodhue [14]	Accuracy, reliability	Currency, level of detail	Compatibility, meaning, presentation, lack of confusion	Accessibility, assistance, ease of use (of h/w, s/w), Locatability
Ballou and	Accuracy,	Completeness,		

Pazer [4]	consistency	timeliness		
Wand and Wang [37]	Correctness, unambiguous	Completeness	Meaningfulness	

Table 2: Classification for academics (Source (Lee et al., 2002, p. 134.))

[a]Classified as system quality rather than information quality by Delone and McLean.

The main difference is that academic approaches try to cover all aspects of IQ, where practitioners focus on particular problems of their context. This separation between academic approaches and practice is interesting, because the MIT group are academics, yet they run the practice-oriented Total Data Quality Management program, which we will discuss shortly.

Note that the aforementioned papers, and others in the tradition, generally do not *define* IQ dimensions, such as objectivity, timeliness, and so on. They primarily *categorise* them. In referring back to the Wang (1998) paper, members of the MIT group talk of having 'empirically derived' quality dimensions. However, note that they generally aim merely to ask practitioners, academics, and information consumers what they take good quality information to be. These surveys certainly make the initial point: information consumers need more than merely accurate information. Yet this point has been made effectively now, and further surveys might best be used to examine more novel aspects of IQ practice. A natural question arises as to what methodology should be used next to help understand IQ in general.

A second methodological approach is adopted in Wand and Wang (1996). The 1996 paper is referred to, but less than other early papers such as Wang (1998). In the paper itself, Wand and Wang refer to it as an 'ontological' approach. Batini and Scannapieco (2006) call it a 'theoretical' approach. We adopt the earlier terminology.

There are various summaries in the paper, but our point is best illustrated by table 4 below, reconstructed from Wand and Wang (1996, p. 94):

DQ Dimension	Mapping Problem	Observed Data Problem
Complete	Certain real world (RW) states cannot be represented	Loss of information about the application domain
Unambiguous	A certain information system (IS) state can be mapped back into several RW states	Insufficient information: the data can be interpreted in more than one way
Meaningful	It is not possible to map the IS state back to a meaningful RW state	It is not possible to interpret the data in a meaningful way
Correct	The IS state may be mapped back into a meaningful state, but the wrong one	The data derived from the IS do not conform to those used to create these data

Table 3: The 'ontological' approach to IQ (source: (Wand & Wang, 1996, p. 94))

Wand and Wang are attempting to understand how IQ errors can be generated. They may also be interested in relations *between* dimensions that surveys may miss. Batini and Scannapieco comment on this paper:

> All deviations from proper representations generate deficiencies. They distinguish between design deficiencies and operation deficiencies. Design deficiencies are of three types: incomplete representation, ambiguous representation, and meaningless states. … Only one type of operation deficiency is identified, in which a state in RW might be mapped to a wrong state in an IS; this is referred to as garbling.' (Batini & Scannapieco, 2006, p. 36.)

Ultimately,

> A set of data quality dimensions are defined by making references to described deficiencies.' (Batini & Scannapieco, 2006, p. 37.)

These dimensions are: complete, unambiguous, meaningful and correct.

Methodologically, the paper is laid out analogously to a mathematical proof, with conclusions apparently derived from axioms or assumptions. In the end, of course, such material can only be analogous to a mathematical proof, and the source of assumptions and the derivations from them are not always clear. Nevertheless, the conclusions are interesting, and it is perhaps better to interpret them as the suggestion of highly experienced academics, who have been thinking about IQ and practising IQ improvement for some time. Then, the test of such conclusions would seem to be whether or not they enhance IQ practice.

Overall, the IQ literature is still seeking a settled method for advancing theoretical understanding of IQ, while even today the field has not fully assimilated the implications of the purpose-dependence of IQ.

1.2 IQ improvement programmes

There have been huge improvements in IQ practice. The MIT group runs what they call a 'Total Data Quality Management' program (TDQM), helping organizations improve their IQ in practice. A further important context for current work has been the development of tools for this programme.

Wang, Allen, Harris, and Madnick (2003, p. 2) summarize the idea of TDQM thus:

> Central to our approach is to manage information as a product with four principles […]:
>
> 1. Understand data consumers' needs,
> 2. Manage information as the product of a well-defined information production process,
> 3. Manage the life cycle of the information product, and
> 4. Appoint information product managers.

Since the 1990s, the focus of TDQM has been to get organizations to ask themselves the right questions, and give them the tools to solve their own IQ problems. The right questions involve understanding the entire process of information in the organization, where it goes and what happens to it, and understanding all the different people who try to use the information, and what they need from it. Then, and only then, can organizations really improve the quality of their information. So the first theme of TDQM is to get information producers to understand, map and control their entire information production process. This is an ongoing task, and TDQM recommends the appointment of information executives on the board of directors of companies, with specific responsibility for managing the company's information flows.

Interwoven with this first theme, the second theme is to allow information consumers to assess for themselves the quality of the information before them, interpret the data semantics more accurately, and resolve data conflicts. This is largely approached using metadata, that is, data about data. Data items are tagged with metadata that allow information users to assess their quality. Such metadata now range widely from an overall IQ score, to something much simpler, such as a source of the data, or an update date and time. This tagging procedure was discussed by Wang, Kon, and Madnick (1993, p. 1):

> In this paper we: (1) establish a set of premises, terms, and definitions for data quality management, and (2) develop a step-by-step methodology for defining and documenting data quality parameters important to users. These quality parameters are used to determine quality indicators, to be tagged to data items, about the data manufacturing process such as data source, creation time, and collection method. Given such tags, and the ability to query over them, users can filter out data having undesirable characteristics.

Here, they are beginning to build the step-procedure that would become central to TDQM. Wang, Reddy, and Gupta (1993, p. 2) write:

> It is not possible to manage data such that they meet the quality requirements of all their consumers. Data quality must be calibrated in a manner that enable consumers to use their own yardsticks to measure the quality of data.

They try to show how to do this for some key dimensions: interpretability, currency, volatility, timeliness, accuracy, completeness and credibility. Wang, Reddy, and Kon (1995, p. 349) are explicit:

> Because users have different criteria for determining the quality of data, we propose tagging data at the cell level with quality indicators, which are objective characteristics of the data and its manufacturing process. Based on these indicators, the user may assess the data's quality for the intended application.

There are some formal problems with this kind of tagging. The most obvious is that of computational power. If you are already struggling to maintain and control a lot of data, tagging every data item with one or more tags quickly multiplies that problem.

Further, one cannot always tag at the cell level – the level of the basic unit of manipulation – as one would prefer. Nevertheless, the idea of the importance of information consumers is being strongly supported in the IQ improvement practice by the use of tagging by metadata aimed at enabling consumers to make their own assessment of information quality.

To achieve this, it is essential to know what an organization does with its information, and what it needs from its information. In a paper where the language of TDQM appears early on, Kovac, Lee, and Pipino (1997, p. 63) write:

> Two key steps are (1) to clearly define what an organization means by data quality and (2) to develop metrics that measure data quality dimensions and that are linked to the organization's goals and objectives.

The whole system must be properly understood to provide real quality information, instead of improving only on a department-by-department, stop-gap basis.

This leads to the development of information product maps (IP-MAP) as an improvement of the earlier 'polygen model' (Wang & Madnick, 1990). Wang (1998) starts using the term 'information product' (IP), and is clearly building the idea of mapping information:

> The characteristics of an IP are defined at two levels. At the higher level, the IP is conceptualized in terms of its functionalities for information consumers. As in defining what constitutes an automobile, it is useful to first focus on the basic functionalities and leave out advanced capabilities (for example, optional features for an automobile such as air conditioning, radio equipment, and cruise control). … Their perceptions of what constitute important IQ dimensions need to be captured and reconciled. (Wang, 1998, p. 61.)

He continues:

> At a lower level, one would identify the IP's basic units and components and their relationships. Defining what constitutes a basic unit for an IP is critical as it dictates the way the IP is produced, utilized and managed. In the client account database, a basic unit would be an ungrouped client account.' (Wang, 1998, p. 63.) In summary: 'The IP definition phase produces two key results: (1) a quality entity-relationship model that defines the IP and its IQ requirements, and (2) an information manufacturing system that describes how the IP is produced, and the interactions among information suppliers (vendors), manufacturers, consumers, and IP managers. (Wang, 1998, p. 63.)

The IP-MAP is developed in more detail, the basic elements of such a map are defined, and the purpose explained:

> Using the IP-MAP, the IP manager can trace the source of a data quality problem in an IP to one or more preceding steps in its manufacture. We define the property of traceability as the ability to identify (trace) a sequence of one or more steps that precede the stage at which a quality problem is detected. Also, given two arbitrary stages in the IP-MAP, we must be able to trace the set of one or more stages, in progressive order, between the two. Using the

> metadata, the individual/role/department that is responsible for that task(s) can be identified and quality-at-source implemented. (Shankaranarayanan, Wang, & Ziad, 2000, p. 15.)

The MIT group have already achieved a great deal in expanding understanding of IQ and IQ practice far beyond simple accuracy measures. This has impacted on all current work. Although they structure their thinking in terms of a business model, we will shortly look at IQ applications in science, and in government, the law and other societal institutions.

1.3 The 'Italian School'

Batini and Scannapieco (2006) is an excellent overview of work on IQ, a presentation of their own work, and a guide to where new work is needed. Batini and Scannapieco are both academics who also practise, and much more of their work – at least the work from which they draw their examples – is work on government-held data, such as address data. They work broadly along the lines of the TDQM programme, but offer extensions to the IP-MAP better to represent the differences between operational processes, using elementary data, and decisional processes using aggregated data, and to track information flows better. They offer a way to compute and represent quality profiles for these. They also offer 'Complete Data Quality Management' (CDQM) which is their improved version of TDQM to take into account the extra resources they have provided. The particular details are not important to this introductory review, and are thoroughly described in Batini and Scannapieco (2006).

Methodologically, Batini and Scannapieco seem to favour what they call the 'intuitive' approach to developing a theoretical understanding of IQ. They write:

> There are three main approaches adopted for proposing comprehensive sets of the dimension definitions, namely, theoretical, empirical, and intuitive. The theoretical approach adopts a formal model in order to define or justify the dimensions. The empirical approach constructs the set of dimensions starting from experiments, interviews, and questionnaires. The intuitive approach simply defines dimensions according to common sense and practical experience.' (Batini & Scannapieco, 2006, p. 36.)

In line with the intuitive approach, Batini and Scannapieco focus firmly on understanding IQ in practice, by allying discussion of dimensions of IQ and their categories with discussion of examples of metrics used to measure those dimensions. They also categorise IQ *activities*. The idea is to examine common things that are done in the process of improving IQ, and understand what the tools and common methods and problems are. They categorise many activities (Batini & Scannapieco, 2006, pp. 70-71), but their aim can be illustrated by looking briefly at the four activities they examine in detail in chapters 4-6.

One very common activity they call 'object identification'. (It is also sometimes called 'record linking', 'record matching', or 'entity resolution'.) This is when you have two

or more sets of data, such as the address data of two different government departments, and you have to identify the records that match the same real-world object – in this case the real house. Data integration is the activity of presenting a unified view of data from multiple, often heterogeneous, sources, such as two sets of address data. Quality composition defines an algebra for composing data quality dimension values. For example, if you have already worked out an IQ score for the completeness of A, and of B, then you need to compute the completeness of the union of A and B. Finally, error localization and correction is the activity performed when the rules on data are known, and you search to find tuples and tables in your data that don't respect the rules, and correct values so that they do. This focus on common activities is a useful practice-oriented way of approaching understanding IQ.

Batini and Scannapieco emphasize that a great deal of work along the lines they have begun is still needed. They write:

> a comprehensive set of metrics allowing an objective assessment of the quality of a database should be defined. Metrics should be related to a given data model or format (e.g., relational, XML, or spreadsheets), to a given dimension (typically a single one), and to different degrees of data granularity. (Batini & Scannapieco, 2006, p. 222.)

No such comprehensive set is available to date. A great deal has been achieved in IQ, and some very good practice has been developed, but much remains to do. Batini and Scannapieco summarise in their preface:

> On the practical side, many data quality software tools are advertised and used in various data-driven applications, such as data warehousing, and to improve the quality of business processes. Frequently, their scope is limited and domain dependent, and it is not clear how to coordinate and finalize their use in data quality processes.
> On the research side, the gap, still present between the need for techniques, methodologies, and tools, and the limited maturity of the area, has led so far to the presence of fragmented and sparse results in the literature, and the absence of a systematic view of the area. (Batini & Scannapieco, 2006, p. IX.)

Thus IQ research has achieved a great deal both in academia and in practice, but still faces significant challenges. The IQ field is vibrant, still finding out what is possible, and facing many challenges with enthusiasm.

2 Developing work

IQ literature and practice is now so sprawling that we cannot hope to offer anything approaching a comprehensive survey of current work. Instead, as a guide, we offer a look at some of the main areas of development, to illustrate the excitement of current work on IQ. Naturally, we focus on issues relevant to the papers in the rest of the book, and we are guided by the conversations we have been privileged enough to have during the course of our project. This makes for an eclectic tour, which illustrates the fascinating diversity of work on IQ.

Data has been growing, but also diversifying. Single databases with well-defined data schemas are no longer the primary problem. Instead, the challenge is to understand and manage different kinds of systems. Peer-to-peer systems do not have a global schema, as peers donating data determine their own schemas, and schema mappings are needed to allow queries across data. On the web, data can be put up in multiple formats, often with no information about provenance. The most important developments are in extending what has already been well understood, in the safer and easier domain of structured data, to the far messier but more exciting domain of unstructured or partially structured data, and to under-examined forms of data, such as visual data.

In this section, we will examine: how work on provenance and trust is applied to assess quality of unstructured data; attempts to build a mid-level understanding to mediate between theory and practice; the extension of well-understood IQ activities, such as object identification, to unstructured data; work on visual data and data visualization; and understanding IQ by understanding error.

The first major area of developing research is IQ in unstructured data, particularly on trust, provenance and reputation. The core idea is very simple: where do the data come from (provenance), are they any good (trust) and is their source any good (reputation)? The approach develops further the idea of the polygen model, which dealt for the first time with the problem of multiple heterogeneous sources. Provenance is generally offered to the user by tagging data with where it comes from, and what has happened to it before it gets to the user. But much more work is needed on how to model and measure the trustworthiness of data and the reputation of particular sources.

An example of work in progress is early research on metrics for trust in scientific data by Matthew Gamble at the University of Manchester.[1] Gamble is working on how scientists trust information from other scientists. This is an interesting correlate of the problem of crowdsourced data: there is equally a problem of the quality of expert-sourced data. The gold standard for most scientists is to be able to reproduce the data – or at least a sample of the data – themselves. But this is often impossible, for reasons of cost, complexity, or simply because of lack of access to necessary technologies. Cost and risk are important, in Gamble's work, as cost and risk frame judgements of good enough quality. If many people are reporting similar results, meaning that they are not very risky, while the results would be costly to reproduce, then further reducing the risk is not worth the high cost. The published results are likely to be trusted (Gamble & Goble, 2011). In this context, Gamble is using provenance traces of data to estimate likely quality of a piece of data. Part of the

[1] We are very grateful to Matthew Gamble for meeting with Phyllis Illari to explain the overview of his project.

provenance given is the experimental technique used to generate the data, although frequently there is information missing, such as average rate of false positives. Trust measures indicators of likely quality, such as the number of citations of a paper. Gamble is borrowing available metrics, and using Bayesian probabilistic networks to represent these metrics in order to calculate likely quality, based on provenance, trust, and so on, currently applied to the likelihood of the correctness of chemical structure. Representing metrics as Bayesian net fragments enables one to join them together, and also to compare them more formally.

In general, the suite of metrics Gamble is developing all have to be adapted to particular situations, but in theory the fragments could be put together with provenance to yield a 'Situation Specific Bayesian Net' to compute an overall quality score of data. In theory, scientists could use it to dump data, or to weight their own Bayesian net according to the quality score of the data. However, this is unlikely in practice. At this stage the work is more likely to yield a benchmark for metrics so that they can be understood and compared in a common way. It also helps to push forward the idea of being able to move from provenance to metrics.

The second area we will look at also explores the connections between theory and domain-specific metrics. Embury and Missier (this volume) explain that work on identifying and categorising dimensions of IQ is no longer proving useful to their practice, and an alternative approach is needed. They developed what they call a 'Quality View' pattern, which is a way of guiding the search for IQ requirements and the information needed for practitioners to create executable IQ measurement components. They survey how they applied this approach in projects involving identifying proteins, in transcriptomics and genomics, and in handling crime data for Greater Manchester Police. The idea is that Quality View patterns guide the application of decision procedures to data. Although they are mid-level between theory and practice, they guide the development of domain-specific metrics appropriate to the particular data in each case. In this way, Embury and Missier, like Gamble, are exploring the space between work on what IQ is, and the development of highly domain-specific metrics.

The third example of developing work is in extending those things we can do well for structured data, in order to figure out how to perform the same tasks for unstructured data. For example, Monica Scannapieco is working on how to extend one of the common IQ activities for structured data – entity matching or record linkage – to unstructured data. Scannapieco calls this 'object matching'. This is the problem of putting together two or more sets of data, when one faces the task of identifying which data in each set refers to the same worldly object. For example, there are many thousands of web pages containing information about cities. How do we decide which pages are all about London, which are about Paris, and so on?

Scannapieco (this volume) examines the problem with respect to two different kinds of relatively unstructured data: linked open data and deep web data. Linked open data are data made available on the web, but linked to related data, most obviously, data about the same real world object – such as data about Paris. For example, DBPedia makes the content of the infoboxes on Wikipedia (the structured part of Wikipedia pages) available in Resource Description Framework (RDF) format, which gives the relationship between items, how they are linked, along with both ends of that link. This is in contrast with what is known as deep web data, which is not directly accessible by search engines, because it consists of web pages dynamically generated in response to particular searches, such as the web page an airline site generates in response to a query about flights on a particular day to a particular destination. Object matching is an issue for both cases, as is the size of the data sets. Scannapieco surveys the issues for addressing object matching, and more general quality issues, in such data. A particular concern is settling on a characterization of identity of two objects.

The fourth example of developing work is work on visualization and visual data. The vast majority of the work on data quality to date has been on the quality of numbers or texts such as names stored in databases, yet presenting data visually is now quite standard. For example, in O'Hara (this volume), maps are used to present crime data to citizens via a website. In Chen, Floridi and Borgo (this volume) the practice of visualisation of data is examined, and the standard story that the purpose of visualisation is to gain insight is questioned. Chen *et al.* argue, by looking at various examples, that the more fundamental purpose of visualization is to save time. Notably, time can be saved on *multiple* tasks that the data are used for, which may of course include gaining insight. In addition to allowing there to be multiple purposes for visualisation, this approach also removes any requirement that it be impossible to perform such tasks without using data visualisation. With these arguments in place, Chen *et al.* argue that the most important metric for measuring the quality of a visualization process or a visual representation is whether it can save the time required for a user or users to accomplish a data handling task.

Batini, Palmonari and Viscusi (this volume) aim to move beyond the much-studied information quality paradigm case of the traditional database, to examine information quality 'in the wild'. They re-examine traditional concepts of information quality in this new realm. In this, they share a great deal with Scannapieco's work, arguing that traditional dimensions, and approaches such as in the ISO standard issued in 2008 (ISO/IEC 25012:2008), still need extensive rethinking. Batini *et al.* study schemaless data by examining the quality of visual data, such as photographs, which are ignored by the ISO standard. They suggest that we can define the quality of an image as the lack of distortion or artefacts that reduce the accessibility of its information contents. Common artefacts are blurriness, graininess, blockiness, lack of contrast and lack of saturation. They note that there are going to be ongoing problems with data quality of, for example, diagrams, as even the most objective-seeming accessibility or

readability guidelines for creating diagrams show cultural specificity. They offer the example that most diagrammers try to have straight lines, with as few crossing lines as possible, but Chinese professors prefer diagrams with crossing and diagonal lines.

The fifth developing area concerns understanding information quality by examining failures in that quality – by better understanding error. This is much as Batini *et al.* do in categorising good images as ones that avoid known classes of problems. This approach has been in the literature at least since Wand and Wang (1996), but it is still being pursued. It is adopted by Primiero (this volume), who sets out to 'define an algorithmic check procedure to identify where a given dimension fails and what kind of errors cause the failure.' (page) Primiero proceeds by applying a broad categorization of errors, in accordance with three main kinds of requirements that can fail when there is error: validity requirements, which are set by the logical and semantic structure of the process; correctness requirements, which are the syntactic conditions for the same process; and physical requirements, which are the contextual conditions in which the information processing occurs. This cross-cuts with three modes of error: conceptual, which relates to configuration and design of the information process; material, or aspects of implementation of the process; and executive, or relating to successful execution of the process. This finally yields four main cases of error (as not all combinations are possible). Primiero uses these to re-examine traditional IQ dimensions such as consistency, accuracy, completeness and accessibility, and assess how failures occur.

Fallis (this volume) uses a similar approach but in a different way. He analyses IQ by classifying various kinds of disinformation – which he takes to be deliberate misinformation. He writes:

> But disinformation is particularly dangerous because it is no accident that people are misled. Disinformation comes from someone who is actively engaged in an attempt to mislead. Thus, developing strategies for dealing with this threat to information quality is particularly pressing.' (page)

Fallis points out that disinformation, unlike a lie, does not have to be a statement but could, instead, be something like a misleading photograph, and disinformation could be true but still designed to mislead by omission. Fallis examines the many different types of disinformation, in an extended attempt to characterize disinformation. He illustrates the variety of kinds of disinformation.

Finally, Stegenga (this volume) illustrates how various approaches to evaluating information quality in medical evidence are attempts to avoid kinds of error. In sum, the attempt to understand error is clearly yielding interesting work, although it may well not yield a unitary approach to information quality, as might have been hoped. This is not surprising if the purpose-dependence of IQ is taken seriously. Just as particular virtues of information are more important for different purposes, so are

particular errors. For some users, late but accurate information is better than speedy but inaccurate information, but not for others.

IQ practice is diversifying, and constantly pushing the boundaries of what is possible. In particular, it is applying existing abilities to unstructured data, such as in understanding the uses and limitations of crowdsourcing, and how to apply techniques that have been developed for structured data in databases to other forms of data such as visual data.

3 Applying IQ

Alongside the deepening theoretical understanding of IQ there have been some extraordinary developments in IQ practice, as information has come to pervade almost all of human activity. For example, the increasing availability of data and its use by multiple people and groups in science means that databases are increasingly crucial infrastructure for science. We refer philosophers in particular to the work of Sabina Leonelli (Leonelli, 2012, 2013; Leonelli & Ankeny, 2012). For data sharing to be effective, data has to be maintained in a form understandable from multiple disciplinary backgrounds, and frequently integrated from multiple sources. So there are extensive applications of the original home of IQ, databases, and newer approaches, such as trust and provenance, in science. The importance of quality information to the well-functioning of society as well is also now hard to underestimate. Frequently, the accessibility of that data to the relevant people is a serious problem, and some data must now be available to all citizens. The two issues of data in science and in society often come together. For example, the absence of longitudinal funding for many scientific databases is a serious impediment in some sciences, and directly impacts society with the handling of medical data (Baker, 2012).

Again, we cannot hope to be comprehensive. We will illustrate the issues of applying IQ by looking at examples of applications to medical data, and to social data.

3.1 Medical data and evidence

There has been a buzz about medical data in recent years, so much so that everyone knows there is a potential problem. But what is interesting on investigation is that there are so many facets of IQ problems in medicine, as it arises in medical discovery, treatment, and maintaining patient records so that patients can be treated appropriately over a lifetime.

One of the core challenges of managing records in healthcare systems is the sheer number of people trying to use the data, and their multiple purposes. Patient records have to be maintained, to be usable by many people with widely varying expertise, including at least family doctors, consultants, nurses, and administrators, and they have

to be kept confidential. What is wanted is an efficient, accessible, easy to update system that can be used without ambiguity by multiple people for multiple purposes. Patient records therefore nicely illustrate how far IQ problems outstrip mere accuracy.[2]

At the moment, such databases are constrained using integrity constraints on what data can be input, which force consistency. First, there are constraints on what data *have* to be input for each patient, such as name, address, sex, age and so on; and text is not usually entered free-text, but from a list of constrained choices. For example, diagnoses of illnesses are coded, and entered by code. Second, there may be constraints across these choices, to weed out errors at the data input stage. For example, a patient cannot be 3 years old and pregnant; or completely healthy and in the intensive care ward.

Such coding systems can be incredibly frustrating for thousands of busy people whose job is not to maintain data, but to care for patients. There is often a separation between the people who use the data, and those who gather it. Those forced to *gather* it may not be as medically informed as those using it, and so struggle to make nuanced choices in difficult to classify cases. Errors are frequent. Further, different *users* of data will maintain data differently. People and institutions are better at maintaining the data that determine what they are paid, and regulated items, such as prescriptions.

Quality assessment is also an issue in the evaluation of medical evidence. First, evidence is assessed for quality when making decisions about effective treatments, and also licensing them to be used, which is done by bodies such as the Food and Drug Administration agency in the US, and the National Institute for Clinical Excellence in the UK. This issue is addressed by Stegenga (this volume). A great deal of work has been done to articulate and standardise methods of assessment of evidence in medicine, particularly by international bodies such as the Cochrane Collaboration (http://www.cochrane.org/). The general idea is to articulate best practice. However, the upshot is often to generate a one-size-fits-all assessment of quality based solely on the method by which the evidence was gathered, without reference to its purpose. Almost all approaches to medical data prioritise evidence produced by Randomised Controlled Trials over other forms of studies. Stegenga examines various Quality Assessment Tools that have been designed and used to assess the quality of evidence reported in particular scientific papers, in an attempt to aggregate evidence and make a decision about the effectiveness of a treatment – and ultimately decide whether it should be licensed. A serious problem with these tools is that different tools often do not agree about the quality of a particular study, and different users of the same tool will often not agree about the quality of a particular study. There are many serious

[2] We thank Andy Bass, Computer Science, Manchester, who works on patient record systems, for personal conversation about these issues.

problems of assessing the quality of medical evidence (Clarke, Gillies, Illari, Russo, & Williamson, 2014; Osimani, 2014).

Information about diseases and effective treatments is available on the web, and patients access it. Further, medical professionals need some way to keep up their expertise once they have finished their formal training, and they also turn to web information. Ghezzi, Chumbers and Brabazon (this volume) describe a variety of measures available to help assess internet sources of medical information. They also describe a course they have designed to train medical students to assess medical evidence on the web, to allow them to update their own expertise, and to talk with patients who may have been misled by what they have read online. Even relatively simple measures, such as checking whether the information comes from a source that is attempting to *sell* the treatment, and searching for references to scientific papers, have proven very effective at weeding out bad information.

IQ is also a problem in the general move to repurpose medical data. Given the expense of data gathering, the ongoing need for more data, and the idea that there are rich resources in data that often go unmined, it is not particularly surprising that there are various moves afoot to make data gathered in one study available for further use. The Food and Drug Administration agency (FDA) in the USA is encouraging this, as is Health Level 7 in Europe. There are significant challenges, as illustrated by the project involving Meredith Nahm, in bioinformatics at Duke[3], which defined the data elements for schizophrenia that the FDA intends to require to be released to the central database before the FDA will license treatments (Nahm, 2012; Nahm, Bonner, Reed, & Howard, 2012). Even with FDA backing for these kinds of projects, trying to get support from experts and funding bodies proved quite a challenge. Ultimately, the project used the DSM-IV, which is the diagnostics manual for psychiatry, and the paperwork generated by clinical professionals, to extract a set of suggested data elements, before engaging in consultation exercises with experts to finalise data elements. However, just before the publication of the updated DSM-V, the NMIH, a major funder of research in psychiatry, announced that it will preferentially fund projects that ignore the DSM categories in favour of their own system. The challenge is that categories of disease and relevant data elements are not settled in psychiatry, or in medicine, and will have to be updated continuously. Projects of this kind will be ongoing.

3.2 Social data

Data have now become a huge concern of society. Again we illustrate the diversity of the impact of information quality on society by examining three cases. First, we look at the quality of personal digital archives. Then we examine the increasingly pressing

[3] We thank Meredith Nahm for discussions.

issue of how to admit only quality information into law courts, given the impossibility of jurors making an informed assessment of such information. Thirdly, we examine the increasing drive to making government data open. This continues the theme of the law, as we will look at crime data, which clearly comes full circle to impact on the private lives of citizens.

First, personal digital archives, such as Facebook timelines, personal and professional files, or family photographs and albums, have become important to people in managing and enjoying their lives. How we disseminate such information, manage its quality, and protect it, is of deep personal and professional concern.

John (this volume) uses the expertise of a professional who manages digital archives for the British Library, to examine the quality of digital archives as they are managed by private individuals.

John lays out seven aspects of quality for digital archives, as a background. But he argues that we should also pay attention to the quality of digital archives 'in the wild' – not only when they enter a specialised repository. This is partly to assist in the job of repositories, but also because the role of personal archives means that their quality affects people's lives. John argues that thinking from an evolutionary perspective – examining how information varies, and is replicated and selected – can help us ask the right questions about quality of personal digital information, and understand better how such information grows, inheriting characteristics of previous archives, such as the growth of a family's archive. This perspective should help, as natural selection has proven good at creating adaptability in the face of uncertainty, which is just what such personal digital archives need. A crucial question for investigation, then, is: are there predictable ways in which digital archives grow in the wild, predictable 'selection pressures' that we can come to understand better, and so better control and compensate for?

The second area we will examine is the quality of expert evidence in the law, specifically in law courts. There is variation across countries, of course, but judges are often asked to ensure that only good quality evidence gets presented in court, and there have been some notable failures. There are currently proposed new rules on expert evidence in the UK. In practice, up until now relatively simple proxy indicators of quality have been favoured, such as the professional qualifications of the expert, membership of professional societies, and peer review and citations of scientific work referenced. Schafer (this volume) discusses how digital media can change this, with particular reference to how digital media can change peer review, which is currently a favoured quality mechanism.

One crucial problem of forensic information being presented in court is the availability of a sensible reference database. The need for such a database to allow estimations of

relevant probabilities came to the fore with DNA, and the situation is much worse for many other kinds of evidence. For example, if you lack a reference database for, say, earprints, then how alike earprints are cannot be estimated accurately, and evidence as to how similar the earprint recovered from the scene is to that of the accused cannot really be given. Schaffer argues that the digital revolution will help with this problem in the future, by allowing access to non-regulated, informal datasets that can allow forensic practitioners to estimate base rates and standards in an unprecedented way.

Schafer also argues that the digital revolution can help with a second problem: the possibility of lawyers and judges assessing whether abstract scientific theories used by experts are 'generally accepted in the scientific community'. Peer review itself cannot indicate whether an idea has come to general acceptance. But digital media are supporting new forms of engagement with science, and allowing access to ongoing discussion of already published papers, including information about post-publication *withdrawal* of papers. Schafer envisages that, in the future, such venues might be routinely data-mined to allow more quantitative assessment of whether research is generally accepted, and suggests that IQ research can help with this task.

The third area we will consider is open data, which O'Hara (this volume) discusses with respect to government data. Open data is made available to anyone who might wish to use it, so it is explicitly presented with no specific user or purpose in mind. This raises similar problems as the repurposing of data in medicine. O'Hara looks at heuristics and institutional approaches to quality in open data, and at how the semantic web might support mechanisms to enhance quality. One idea associated with open data is that increased scrutiny will improve the quality of data, by detecting errors, leading to the idea of crowdsourced data improvement.

O'Hara discusses a particular initiative to make local crime data available to citizens in the UK, to allow them to take it into account in decisions such as where to live, and routes to travel. The project met problems integrating data from 43 different police forces in the UK, lacking any national geodata coding standard. Further, burglaries and assaults have a definite location that can be mapped, but this is not true of all crimes, such as identity theft. It was also difficult to maintain anonymity. If a burglary is shown as taking place at your address, then you are identified as the victim of that crime, perhaps against your wishes. Reasonable data accuracy was reconciled with the need for some anonymity by making the data available on location vaguer, giving number of crimes by small geographical area, rather than a precise location for each one. O'Hara suggests that data producers designing such a system should interact with likely users to make the data accessible. The decision was to compensate for problems in the data by making users as aware as possible of the possible limits of the data they were given, using metadata. So note that in the end getting such open data systems to work is difficult without *some* attention to possible users of the information.

Ultimately, then, these three cases illustrate how pervasive information quality issues are, and how they impact on the daily lives of everyone in society.

4 Conclusion: Theoretical challenges

The concluding two papers of the book finish where we started, as Illari and Floridi examine the theoretical problem of purpose-dependence of IQ, as pressed by the MIT group. Illari (this volume) takes up purpose-dependence alongside the practical problem that successful metrics for measuring IQ are highly domain-specific and cannot be transferred easily. She argues that both theoretical and practical approaches to IQ need to be framed in terms of an understanding of these deep problems. She supports a categorisation of IQ dimensions and metrics that highlights, rather than obscures, these problems.

Floridi (this volume) examines purpose-dependence alongside the argument that the problem of big data is often not the amount of data, but the difficulty of the detection of small patterns in that data. IQ concerns the possibility of detecting these patterns. Floridi argues for a 'bi-categorical' approach to IQ that allows it to be linked explicitly to purpose.

These issues play out in many of the papers in the volume. Purpose-dependence inhibits the possibility of inter-level theorising about IQ, creating understanding that lies between what IQ is, dimension categorisations, and domain-specific metrics. This is addressed by the Embury and Missier paper (this volume) and in the work by Gamble that we have discussed, and shows the importance of this work.

This background also illuminates the attempt to address IQ comprehensively by categorising error, shared in this volume by Primiero, Fallis and in some ways by Batini *et al.* and Stegenga. This approach is undeniably valuable, but a comprehensive assessment may be too much to hope for. It is likely that different kinds of error are more or less important for different purposes.

In medical evidence, discussed by Stegenga (this volume), we see the impact of pursuing an ideal of a purpose-independent estimation of quality of evidence. The way traditional evidence assessments proceed, quality of evidence is ideally independent of *everything* except the method used to generate the evidence. Against the background of this IQ literature, the deep difficulties with such an approach are clear.

The scale of data now available in medical research also underlines the small patterns problem. Increasingly, our ability to process data – to find the small patterns we seek – is the critical problem. Purpose rules here, too. More data is no good if it merely obscures the pattern you are looking for in your dataset. There needs to be more

attention explicitly to discriminating amongst purposes in assessing fitness for purpose, allowing us better to recognise which data is worth holding on to.

This is an interesting backdrop to the moves to assess information quality in the wild, which we find here in both Batini *et al.*, and John. Learning to deal with information in its natural form, and extract what we need from it there, should help address this problem. This is aligned, then, with work on dealing with unstructured data, such as examining object matching (Scannapieco, this volume), and making data open partly to allow increased scrutiny (O-Hara, this volume).

In short, IQ is a challenging and exciting area of research, already bearing fruit, and certain to reward further research.

References

Baker, M. (2012). Databases fight funding cuts. *Nature, 489*(19). doi: 10.1038/489019a

Batini, C., & Scannapieco, M. (2006). *Data quality: Concepts, methodologies and techniques*. Berlin; New York: Springer.

Clarke, B., Gillies, D., Illari, P., Russo, F., & Williamson, J. (2014). Mechanisms and the evidence hierarchy. *Topoi*. doi: 10.1007/s11245-013-9220-9

Gamble, M., & Goble, C. (2011, June 14-17 2011). *Quality trust and utility of scientific data on the web: Towards a joint model.* Paper presented at the WebSci'11, Koblenz, Germany.

Kovac, R., Lee, Y. W., & Pipino, L. L. (1997). *Total data quality management: The case of IRI.* Paper presented at the Conference on Information Quality, Cambridge, MA.

Lee, Y. W., Strong, D. M., Kahn, B. K., & Wang, R. Y. (2002). AIMQ: A methodology for information quality assessment. *Information & Management, 40*(2), 133-146. doi: 10.1016/s0378-7206(02)00043-5

Leonelli, S. (2012). Classificatory theory in data-intensive science: The case of open biomedical ontologies. *International Studies in the Philosophy of Science, 26*(1), 47-65.

Leonelli, S. (2013). Integrating data to acquire new knowledge: Three modes of integration in plant science. *Studies in the History and the Philosophy of the Biological and Biomedical Sciences: Part C, 4*(4), 503-514.

Leonelli, S., & Ankeny, R. (2012). Re-thinking organisms: The epistemic impact of databases on model organism biology. *Studies in the History and Philosophy of the Biological and Biomedical Sciences, 43*, 29-36.

Nahm, M. (2012). *Knowledge acquisition from and semantic variability in schizophrenia clinical trial data.* Paper presented at the ICIQ 2012, Paris.

Nahm, M., Bonner, J., Reed, P. L., & Howard, K. (2012). *Determinants of accuracy in the context of clinical study data.* Paper presented at the ICIQ 2012, Paris.

Osimani, B. (2014). Hunting side effects and explaining them: Should we reverse evidence hierarchies upside down? *Topoi*. doi: 10.1007/s11245-013-9194-7

Shankaranarayanan, G., Wang, R. Y., & Ziad, M. (2000). *IP-Map: Representing the manufacture of an information product.* Paper presented at the 2000 Conference on Information Quality, MIT.

Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. [Article]. *Communications of the ACM, 39*(11), 86-95. doi: 10.1145/240455.240479

Wang, R. Y. (1998). A product perspective on total data quality management. [Article]. *Communications of the ACM, 41*(2), 58-65. doi: 10.1145/269012.269022

Wang, R. Y., Allen, R., Harris, W., & Madnick, S. E. (2003). *An information product approach for total information awareness.* Paper presented at the IEEE Aerospace Conference.

Wang, R. Y., Kon, H. B., & Madnick, S. E. (1993). *Data quality requirements analysis and modelling.* Paper presented at the Ninth International Conference of Data Engineering Vienna.

Wang, R. Y., & Madnick, S. E. (1990). *A polygen model for heterogeneous database-systems: The source tagging perspective*.

Wang, R. Y., Reddy, M. P., & Gupta, A. (1993). *An object-oriented implementation of quality data products.* Paper presented at the WITS-'93, Orlando, Florida.

Wang, R. Y., Reddy, M. P., & Kon, H. B. (1995). Toward quality data: An attribute-based approach. *Decision Support Systems, 13*(3-4), 349-372. doi: http://dx.doi.org/10.1016/0167-9236(93)E0050-N