Carlo Batini, Monica Scannapieco

Data and Information Quality

SPIN Springer's internal project number, if known

– Monograph –

June 18, 2015

Springer

Chapter 1
Quality of Web Data and Quality of Big Data: Open Problems

Monica Scannapieco and *Laure Berti*

1.1 Introduction

In this chapter we discuss some open issues related to two typologies of information sources that nowadays are particularly significant, namely: Web data and Big Data.

Searching and using the information stored on billions of Web pages poses significant challenges, because this information and related semantics are usually more complex and dynamic than the information that traditional database management systems store [306, 58]. As an evolving collection of inter-related files on one or more Web servers, Web data is extremely rich and diverse, combining multiple types of media and data.

The vision of the Semantic Web aims to make use of semantic representations on the Web at the largest possible scale. Large knowledge bases such as DBpedia (`http://dbpedia.org/`), GovTrack (`http://www.govtrack.us/`) and OpenCyc (`http://www.cyc.com/platform/opencyc`) are freely available as Linked data and SPARQL endpoints, see Chapter ?? for a systematic introduction to Linked data. However, users of such large Semantic Web knowledge bases are often facing three important problems:

- Limited access, due to the lack of high-quality keyword-based searches and the lack of deep-Web access [424, 58]: users can hardly know which identifiers are used and are available for the construction of their queries. Furthermore, domain experts might not be able to express their queries in a structured form, although they have a very precise idea of what kind of results they would like to retrieve [599, 598, 678];
- Limited knowledge of the various IQ problems existing in the Web data: for example, data extracted from semi-structured or even unstructured sources, such as DBpedia or Yahoo Finance (`http://finance.yahoo.com/`) often contain inconsistencies as well as misrepresented, redundant, obsolete, inaccurate or incomplete information [402] the users may even not be aware of. They usually do not have appropriate tools to evaluate, control or monitor IQ.
- Dissatisfaction and misuse of the available Web data or services: depending on the level of quality required, retrieved data or accessed services may not fit for the intended use. For example, in Wikipedia, it may happen, though

in relatively few cases, that some information or some facts are missing or incomplete for general information purposes. But for self-medication, the same quality level may be completely insufficient. To date, some Web-based e-commerce service systems, such as `amazon.com` and `expedia.com`, register every user's past traversal or purchase history and build customer profiles from that data. Based on a user's profile and preferences, these sites select appropriate sales promotions and recommendations, thereby providing better quality of service than sites that do not track and store this information. Although a personalized Web service based on a user's traversal history could help recommend appropriate services, a system usually cannot collect enough information about a particular individual to warrant high-quality recommendations [497].

As a consequence, one of the most important challenges is to determine the quality of Web data – created with HTML and XML, or generated dynamically by underlying Web database service engines – and make this quality information fully and relevantly usable and exploitable.

Two relevant paradigms for characterizing the quality of Web data are trustworthiness and provenance. Trustworthiness can be characterized on the basis of three dimensions, namely: *believability*, *verifiability* and *reputation*. Provenance is a rather complex concept that has been being investigated since several years, and that recently, with the advent of the Web of data, has become even more important. In Section 1.2 we will characterize and describe both paradigms, as well as relevant tools and techniques for dealing with them.

Ensuring IQ is obviously a substantial challenge in Web data management as it involves a set of autonomously evolving data sources that need to be monitored and possibly cleaned for data integration. To the purpose a very relevant task is object identification, that aims at identifying pairs of data-objects that represent the same real world object, and that has been discussed in the context of well grounded types of data in Chapter **??** and Chpater **??**. When objects are Web data, some relevant features must be taken into account, namely: Web data can be highly time-dependant and their quality must be assessed. In Section 1.3 we will describe some relevant issues related to object identification of Web data and we will also describe possible techniques solving such issues.

The second part of the chapter will deal with Big Data. In Section 1.4 after a general characterization of Big Data sources, we will describe issues in characterizing the quality of such data. In particular, we will outline that Big Data involves very different types of sources, and hence their quality characterization does need to be source-specific. In this direction, an overview of approaches for quality characterization of sensor data will be presented in Section 1.5 as an example of how a source-specific quality characterization of Big Data can be carried out.

Specification of Big Data quality can also be dependent on specific application domains, i.e. it is domain-specific. In this respect, in Section 1.6 we will provide an example of quality issues in dealing with Big Data that come from the Official Statistics domain.

1.2 Two Relevant Paradigms for Web Data Quality: Trustworthiness and Provenance

1.2.1 Trustworthiness

In the following we first define concepts related to trust, and then we discuss their interrelationships.

Trust is a level of subjective and local probability with which an agent assesses that another agent will perform a particular action. *Trustworthiness* is the objective probability that the trustee performs a particular action on which the interests of the truster depend. In other words, trustworthiness is the assurance that a system will perform as expected. Though trust and trustworthiness are two distinct concepts, when dealing with techniques for assessing them, the two concepts play often a single role; hence in the following the two terms will be used interchangeability unless needing specific characterizations.

Thirunarayan et al. [608] provide a comprehensive ontology to capture trust-related concepts as well as a detailed comparative analysis of trust models and metrics in diverse contexts. They classify the approaches into *direct trust* referring to trust determined using firsthand experiences over a period of time and *indirect trust* referring to trust determined using experiences of others via referrals. They describe and compare Bayesian approaches to direct trust and trustworthiness in reputation-based processes.

Recent work on performing trust analysis based on the data provided by multiple sources has been proposed. Yin et al. [687] introduced a heuristic fact-finder algorithm *TruthFinder* which performs trust analysis on a providers' facts network. This work was followed by various fact finder algorithms in the context of trust propagation [633] and truth discovery analysis [183]. The goal is the find the truth about some questions or facts given multiple conflicting sources. The proposed approaches differ in the factors taken into account to estimate *source accuracy* and *trustworthiness* e.g., the difficulty of the questions [255], the type of errors [693], the applications (Wikipedia or collection of Web documents) [633] or some potential dependence (through copying relationships) between sources [183].

According to [633], a general approach to trust assessment uses (1) domain-dependent properties for determining trustworthiness based on content and on external metadata, and (2) domain-independent mapping to trust levels through quantification and classification. For example, Wikipedia articles can be assessed based on domain-dependent content-based quality factors such as references to peer-reviewed publications, proportion of paragraphs with citation, article size, and also metadata-based credibility factors such as author connectivity, edit pattern and development history, revision count, proportion of reverted edits, mean time between edits and mean edit length. Another example is the estimation of a website's trustworthiness based on the levels of sensitivity of exchanged information with highly trusted sites (e.g., for identity and banking information exchanges).

1.2.1.1 Trustworthiness Dimensions

There are three dimensions for characterizing trustworthiness, namely: *believability*, *verifiability* and *reputation* that are displayed along with their respective metrics in Table 1.1. The reference for each metric is provided in the table.

In the following, a detailed characterization of the three dimensions is provided.

1.2.1.2 Believability

Believability refers to the extent to which information is regarded as true and credible. Believability can also be defined as the subjective measure of a users belief that the data is "true" [337]. Believability is measured as follows (see also Table 1.1):

- compute the trustworthiness of RDF statements based on provenance information and on the opinion of other information consumers: the system applies a trust function which assigns a trust value which can be a value in the interval [-1,1] where 1: absolute belief, -1: absolute disbelief and 0:lack of belief/disbelief. The trust functions that computes a trust value are based on user-based ratings, provenance-based or opinion-based method.
- compute the trustworthiness of an entity, namely an object or a resource: an objective trust measure for each entity is provided a priory by a trusted third party which provides information such as citation count or global reputation. Once each entity has been given its trust value then it is possible to make trust inference on the new arriving entities.
- compute the trust between two entities by using a combination of (1) a propagation algorithm which utilises statistical techniques for computing trust values between two entities through a path and (2) an aggregation algorithm based on a weighting mechanism for calculating the aggregate value of trust over all paths.
- acquiring content trust from users: based on associations that transfer trust from entities to resources.

Dimension	**SubDimension**	**Description**
Believability	computing the trustworthiness of RDF statements	computing a trust value based on user-based ratings or opinion-based method [299]
	computing the trust of an entity	construction of decision networks informed by provenance graphs [257]
	accuracy of computing the trust between two entities	by using a combination of (1) propagation algorithm which utilizes statistical techniques for computing trust values between 2 entities through a path and (2) an aggregation algorithm based on a weighting mechanism for calculating the aggregate value of trust over all paths [568]
	acquiring content trust from users	based on associations that transfer trust from entities to resources [264]
	detection of trustworthiness, reliability and credibility of a data source	use of trust annotations made by several individuals to derive an assessment of the sources' trustworthiness, reliability and credibility [265]
	assigning trust values to data/sources/rules	use of trust ontologies that assign content-based or metadata-based trust values that can be transferred from known to unknown data [337]
	determining trust value for data	using annotations for data such as (i) blacklisting, (ii) authoritativeness and (iii) ranking and using reasoning to incorporate trust values to the data [85]
	meta-information about the identity of information provider	checking whether the provider/contributor is contained in a list of trusted providers [79]
Verifiability	verifying publisher information	stating the author and his contributors, the publisher of the data and its sources [242]
	verifying authenticity of the dataset	whether the dataset uses a provenance vocabulary, eg. the use of the Provenance Vocabulary [242]
	verifying correctness of the dataset	with the help of unbiased trusted third party [79]
	verifying usage of digital signatures	signing a document containing an RDF serialisation or signing an RDF graph [242]
Reputation	reputation of the publisher	survey in a community questioned about other members [264]
	reputation of the dataset	analyzing references or page rank or by assigning a reputation score to the dataset [437]

Table 1.1 Comprehensive list of IQ metrics for trust dimensions

- detection of trustworthiness, reliability and credibility of a data source: use of trust annotations made by several individuals to derive an assessment of the sources' trustworthiness, reliability and credibility.
- assigning trust values to data sources/rules: use of trust ontologies that assign content-based or metadata-based trust values that can be transferred from known to unknown data.
- determining trust value for data: using annotations for data such as (i) blacklisting, (ii) authoritativeness and (iii) links-based ranking.
- meta-information about the identity of information provider: checking whether the provider/contributor is contained in a list of trusted providers.

Another method proposed by Tim Berners-Lee was that Web browsers should be enhanced with an "Oh, yeah?" button to support the user in assessing the believability of data encountered on the web[1]. Pressing of such a button for any piece of data or an entire dataset would contribute towards assessing the believability of the dataset.

According to the last three points in the listing given before we can point out that believability is measured by checking whether the contributor is contained in a list of trusted providers. There exists an interdependency between the data provider and the data itself. On the one hand, data is likely to be accepted as true if it is provided by a trustworthy provider. On the other hand, the data provider is trustworthy if it provides true data.

1.2.1.3 Verifiability

Verifiability refers to the degree by which a data consumer can assess the correctness of a dataset.

Verifiability is described as the "degree and ease with which the information can be checked for correctness" [79]. Similarly, in [242] the verifiability criterion is used as the means a consumer is provided with, which can be used to examine the data for correctness. Without such means, the assurance of the correctness of the data would come from the consumer's trust in that source. It can be observed here that on the one hand the authors in [79] provide a formal definition whereas the author in [242] describes the dimension by providing its advantages and metrics.

Verifiability can be measured either by an unbiased third party, if the dataset itself points to the source or by the presence of a digital signature (see Table 1.1).

As an example, if we assume that a flight search engine crawls information from arbitrary airline websites, which publish flight information according to a standard vocabulary, there is a risk for receiving incorrect information from malicious websites. For instance, such a website publishes cheap flights just to attract a large number of visitors. In that case, the use of digital

[1] `http://www.w3.org/DesignIssues/UI.html`

signatures for published RDF data could allow to restrict crawling only to verified datasets.

Verifiability is an important dimension when a dataset includes sources with low believability or reputation. This dimension allows data consumers to decide whether to accept provided information. One means of verification in linked data is to provide basic provenance information along with the dataset, such as using existing vocabularies like SIOC, Dublin Core, Provenance Vocabulary, the OPMV[2] or the recently introduced PROV vocabulary[3]. Yet another mechanism is the usage of digital signatures [113], whereby a source can sign either a document containing an RDF serialisation or an RDF graph. Using a digital signature, the data source can vouch for all possible serializations that can result from the graph thus ensuring the user that the data she receives is in fact the data that the source has vouched for.

1.2.1.4 Reputation

Reputation is a judgment made by a user to determine the integrity of a source. It can be associated with a data publisher, a person, organization, group of people or community of practice or it can be a characteristic of a dataset (see Table 1.1).

The authors in [264] associate reputation of an entity (i.e. a publisher or a dataset) either as a result from direct experience or recommendations from others. They propose the tracking of reputation through a centralized authority or, in alternative, via decentralized voting.

Reputation is usually a score, for example, a real value between 0 (low) and 1 (high). There are different possibilities to determine reputation and can be classified into human-based or (semi-)automated approaches. The human-based approach is via a survey in a community or by questioning other members who can help to determine the reputation of a source or by the person who published a dataset. The (semi-)automated approach can be performed by the use of external links or page ranks.

The provision of information on the reputation of data sources allows conflict resolution. For instance, several data sources report conflicting prices (or times) for a particular flight number. In that case, a search engine can decide to trust only the source with higher reputation.

Reputation is a social notion of trust [270]. Trust is often represented in a web of trust, where nodes are entities and edges are the trust values based on a metric that reflects the reputation one entity assigns to another [264]. Based on the information presented to a user, she forms an opinion or makes a judgement about the reputation of the dataset or the publisher and the reliability of the statements.

---

[2] `http://open-biomed.sourceforge.net/opmv/ns.html`

[3] `http://www.w3.org/TR/prov-o/`

1.2.2 Provenance

Representing and analyzing provenance is a topic of research since a decade [300, 601]. Bunemann *et al.* [99] identify several open issues for data provenance of Web data such as: *i)* obtaining provenance information, *ii)* citing components of a data resource that may be (components of) another resource in another context, and *iii)* ensuring integrity of citations under the assumption that cited data resources evolve.

Not knowing the exact provenance used to produce a published dataset often renders the dataset useless (and not only from a scientific point of view). While there has been substantial work on database and workflow provenance, the two problems have generally been examined in isolation. Database provenance, that has been investigated in Chapter **??** is fine-grained and captures precise – why, where and how – dependencies [129] between data and queries. These dependencies are used to formally analyze and improve the quality of data and query results. In contrast, workflow provenance is represented at a coarser level and reflects the functional model of workflow systems which is stateless (each computational step derives a new artifact). Workflow provenance is mainly used to achieve reproducibility of workflow executions.

On the positive side, capturing provenance information is facilitated by the widespread use of workflow tools for processing scientific data and more recently open data. The workflow process describes all the steps involved in producing a given dataset, and hence captures its lineage. Efficiently deriving [19, 88] storing and querying [22] provenance information is still an important research issue in both database and workflow environments.

For Web data, we are confronted with several challenging issues concerning provenance information management. The first challenge addresses the problem of keeping track of Web data lineage from its origin to its final uses and consists in defining and implementing tools for capturing and querying provenance information of data-centric workflows. These tools have to combine database and workflow provenance techniques [133] that specialize general data-oriented transformations, such as the ones specified for warehousing systems [159].

The second challenging issue addresses the problem of building and increasing confidence in the data and consists in using provenance information for capturing and improving the quality of data manipulated by SPARQL queries [178]. The goal is to define appropriate abstract provenance models that capture the relationship between query results and source data by taking into account the employed query operators [606].

With the development of the Linked Data initiative [408], the provenance of that data becomes an important factor for developing new Semantic Web applications. A dedicated W3C group, the Provenance Working Group, part of the W3C Semantic Web Activity [635], developed a set of documents, collectively named as the PROV Family of Documents [634], with the purpose of promoting and enabling representation and interchange of provenance information using widely available formats such as RDF and XML. The following section describes some interesting outcomes of this W3C standardization activity.

1.2.2.1 Provenance on the Web

Provenance of a resource is a record of metadata containing descriptions of the entities and activities involved in producing and delivering or otherwise influencing a given object. The main usage of provenance are related to: (i) understanding where data come from, (ii) identifying ownership and rights over a resource, (iii) making judgments about a resource to determine whether to trust it, (iv) verifying that the process used to obtain a result complies with given requirements, and reproducing it.

Three different perspectives on provenance can be considered:

- *Agent-centered provenance*, that is, what people or organizations were involved in generating or manipulating a resource. For example, in the provenance of a picture in a news article, it is possible to capture the photographer who took it, the person that edited it, and the newspaper that published it.
- *Object-centered provenance*, by tracing the origins of portions of an entity, i.e. an object or a resource, to other entities.
- *Process-centered provenance*, capturing the activities and steps taken to generate a resource. For example, some statistical data are the result of a data collection phase that involved a certain sample, of a data correction phase that involved specific imputation techniques and of a data estimation phase, performed according to defined methods.

The relationships among the different perspectives are shown in Figure 1.1. Key dimensions concerning provenance are shown in Table 1.2, and are: *content*, *management* and *use*. The four dimensions characterizing the content dimension aim to take into account who provided the content (*attribution*), how the content was generated (*process*), how it evolved in time (*evolution* and *versioning*), notes on the content (*justification for decision*) and content it was derived from (*entailment*). Management is instead described by the availability of provenance information (*publication*) as well as its accessibility (*access*), and by non-functional provenance requirements like control policies (*dissemination control*) and performance (*scale*). Finally, the use dimension is characterized by usability aspects (*understanding*), integration aspects (*interoperability* and *comparison*), provenance verifiability (*accountability* and *trust*) and error management issues (*imperfection* and *debugging)*.

The PROV Family of documents collectively consists of eleven documents. Each document can be classified according to the specific type of audience it is intended for, namely:

Category	Dimension	Description
Content	Attribution	Provenance as the sources or entities that were used to create a new result *Responsibility*: Knowing who endorses a particular piece of information or result *Origin*: recorded vs. reconstructed, verified vs. non-verified, asserted vs. inferred
	Process	Provenance as the process that yielded an artifact *Reproducibility* (e.g. workflows, mashups, text extraction) *Data Access* (e.g. access time, accessed server, party responsible for accessed server)
	Evolution and versioning	*Republishing* (e.g. re-tweeting, re-blogging) *Updates* (e.g. a document with content from various sources and that changes over time)
	Justification for decisions	Includes argumentation, hypotheses, why-not questions
	Entailment	Given the results to a particular query, what axioms or tuples led to those result
Management	Publication	Making provenance information available (expose, distribute)
	Access	Finding and querying provenance information
	Dissemination Control	Track policies specified by creator for when/how an artifact can be used *Access Control*: incorporate access control policies to access provenance information *Licensing*: stating what rights the object creators and users have based on provenance *Law enforcement*(e.g. enforcing privacy policies on the use of personal information)
	Scale	how to operate with large amounts of provenance information
Use	Understanding	End user consumption of provenance *Abstraction*: multiple levels of description, summary *Presentation, Visualization*
	Interoperability	Combining provenance produced by multiple different systems
	Comparison	Finding what is in common in the provenance of two or more entities (e.g. two experimental results)
	Accountability	The ability to check the provenance of an object with respect to some expectation *Verification* of a set of requirements *Compliance* with a set of policies
	Trust	Making trust judgments based on provenance *Information quality* *Reputation, Reliability*
	Imperfections	Reasoning about provenance information that is not complete or correct *Incomplete provenance* *Uncertain,probabilistic provenance* *Erroneous provenance* *Fraudulent provenance*
	Debugging	Using provenance to detect bugs or failures of processes.

Table 1.2 Dimensions of Provenance of Web Data

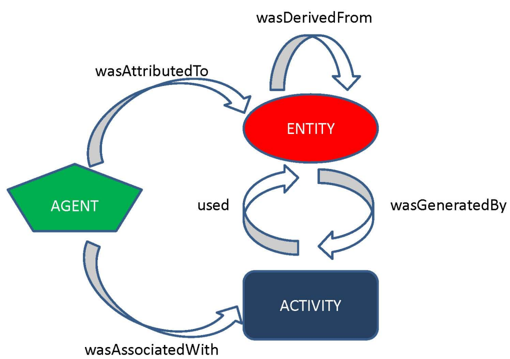

Fig. 1.1 Key concepts of the PROV Family of Documents

- users, that want to understand PROV and use applications that support PROV;
- developers, that want to develop or build applications that create and consume provenance using PROV;
- advanced, that want to create validators, new PROV serializations, or other advanced provenance-based systems.

Table 1.3 lists the PROV framework documents according to the different types of users, namely *users*, *developers* and *advanced*. While the set of documents related to the developers view is of immediate practical usage for provenance publishers, the set of documents that is apart of the advanced view is more intended to be used for both (i) formal definition of the framework's concepts and (ii) provision of specifications for developers of tools that can support provenance publication and validation.

Among the documents shown in Table 1.3, it is particularly relevant the PROV-O document that defines an OWL2 ontology enabling the representation of provenance information for Linked open data. In this respect, it provides both a data model and a technical solution to associate provenance information to Linked open data.

1.3 Web Object Identification

We have seen in Chapter **??** and Chapter **??** that the scope of object identification (OID) is very huge, going from structured data to images (image

Audience	Document Name	Description
Users	Prov-Primer	It is the entry point to PROV offering an introduction to the provenance data model. This is where you should start and for many may be the only document needed.
Developers	Prov-O	It defines a light-weight OWL2 ontology for the provenance data model. This is intended for the Linked Data and Semantic Web community.
	Prov-XML	Defines an XML schema for the provenance data model. This is intended for developers who need a native XML serialization of the PROV data model.
	Prov-AQ	Defines how to use Web-based mechanisms to locate and retrieve provenance information.
	Prov-DC	Defines a mapping between Dublin Core and PROV-O
	Prov-Dictionary	Defines constructs for expressing the provenance of dictionary style data structures.
Advanced	Prov-DM	It defines a conceptual data model for provenance including UML diagrams. PROV-O, PROV-XML and PROV-N are serializations of this conceptual model.
	Prov-N	Defines a human-readable notation for the provenance model. This is used to provide examples within the conceptual model as well as used in the definition of PROV-CONSTRAINTS.
	Prov-CONSTRAINTS	Defines a set of constraints on the PROV data model that specifies a notion of valid provenance. It is specifically aimed at the implementors of validators.
	Prov-Sem	Defines a declarative specification in terms of first-order logic of the PROV data model.
	Prov-LINKS	Defines extensions to PROV to enable linking provenance information across bundles of provenance descriptions.

Table 1.3 Dimensions of Provenance of Web Data

matching) and to completely unstructured information like documents (document matching).

The focus of this section is restricted to Web data, which is a huge category as well. From an OID perspective, Web data can be characterized by some relevant features, listed in the following:

- time variability, considering the time dependency of most of Web data;
- quality, in terms of its characterizing IQ dimensions.

In the following, for each of the above listed features, we will illustrate the impact on the OID process, and some examples of research works that address the OID problem with respect to the specific feature under analysis.

1.3.1 Object Identification and Time Variability

In Chapter **??**, we introduced the concept of time variability of data and of its impact on data and information quality. When considering specifically Web data, the relationship with time has two main aspects. The first one is data volatility, i.e. a temporal variability of the information the data are meant to represent: there are data that are highly volatile (e.g. stock options), other which exhibit some degree of volatility (e.g. product prices), and some which are not volatile at all (e.g. birth dates). The second aspect is more generally related to the time features of the data generating mechanism. For instance, some Web data spring up and get updated in an almost unpredictable fashion, so that their time dimension is not available in a direct way, but does need to be re constructed, if wishing to use those data in any meaningful analysis.

1.3.1.1 Need for fully automated methods

From an OID perspective, the data volatility aspect has the direct implication that manual tasks are not anymore possible (or at least are hard to be executed) during the OID process, that is the process should be fully automated. Decision models for OID are often supervised or semi-supervised, or, in other words, selected record pairs (typically the more difficult to classify) are sent to be clerically reviewed and training set of prelabeled record pairs can be prepared. Implementations of the Fellegi and Sunter model [233] (see also Chapter **??** for an introduction to the model) are often classified as unsupervised methods for learning the status of matching or non-matching of object pairs. However, such implementations are not actually fully automated, as it would be necessary in a OID process on Web data. As an example, several implementations of Fellegi and Sunter rely on the Expectation Maximization (EM) algorithm [176] for the estimation of the parameters of the model. However, in these techniques, manual intervention is required due to *(i)* the need of setting thresholds for identifying matching and non-matching pairs; *(ii)* possible unsuccessful parameter estimation via the EM algorithm (that may happen for instance if the size of the search space is too huge or too much limited).

An example of fully automated technique that can fit the fully automation requirement of Web data is provided in [689], where a statistical approach based on *mixture models* is adopted. More specifically, OID methods rely on distance (or similarity) measures between objects pairs. Due to the stochastic nature of every real-world data generating process, such pairwise distances can be seen as (realizations of) a random variable. Thus, the intuition behind the use of mixture models is that the observed distances arise from a superposition of two distinct probability distributions: the one stemming from the subpopulation of matches and the other from that of non-matches. The ultimate aim of this statistical perspective is to exploit the mixture model for classification purposes, i.e., to bring to light the hidden grouping of the pairs in the underlying M and U classes. To such a scope, the distance is viewed as an observable auxiliary random variable that can be used to make inference on a latent interest random variable, namely the class-membership

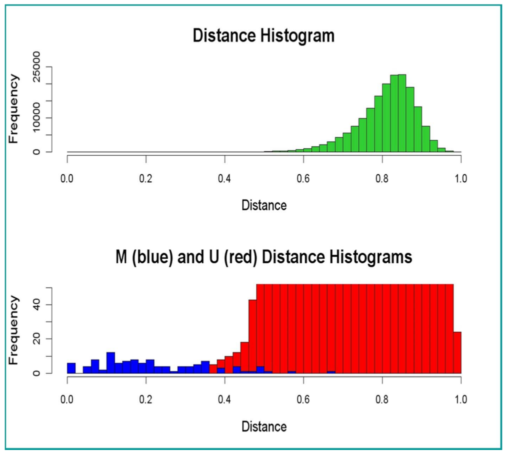

Fig. 1.2 Distance histogram and matches and non-matches histograms: the non-matches red histogram in the lower panel has been cut to allow the detection of the very small matches distribution (blue)

indicator of the pairs. The whole picture is founded upon the hypothesis that the probability distribution of the distance is significantly different inside the M and U classes (see also Chapter **??**, where Section **??** focuses on the impact that the different distributions has on OID metrics). Luckily this is almost always the case in real application scenarios, because typically errors affect data at moderate rates. Whenever such condition holds, the shapes of the M and U distance densities are indeed very different: (i) non-matches tend to be concentrated at higher distances than matches, which furthermore generally exhibit their own distinctive peak at zero distance; (ii) M and U densities show only a relatively small overlap. These qualitative features are so general that one can rightly consider them as a piece of prior knowledge about the underlying (unknown) M and U distance probability distributions: we refer to it as PK1.

Besides PK1, another piece of prior knowledge is readily available in OID applications, namely that matches are rare as compared to Unmatches. We refer to this second kind of prior knowledge as PK2. In Figure 1.2, the distribution of distances for a real dataset (the Restaurant dataset of the Riddle online repository) is shown. In [690], a system called MAERLIN (the acronym stands for Mixture-based Automated Effective Record LINkage) implements the novel suite of methods proposed in [689], and shows how exploiting PK1

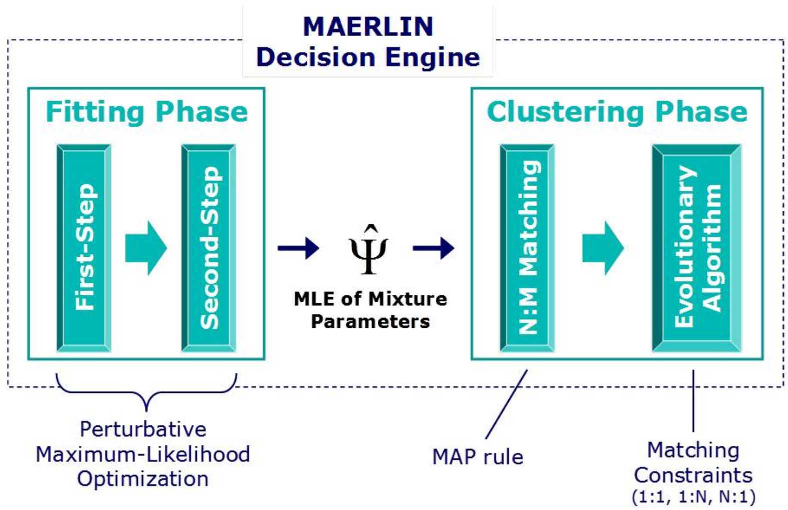

Fig. 1.3 MAERLIN decision engine

and PK2 when facing practical OID tasks. MAERLIN represents the probability density function of the distance as a two-component Beta mixture. The system structures the decision phase of an OID process into two consecutive tasks, as schematically depicted in Figure 1.3. First, it finds (constrained) maximum-likelihood estimates for the mixture parameters by fitting the model to the observed distance measures between pairs. Then, it obtains a probabilistic clustering of the pairs into matches and non-matches by exploiting the fitted model.

The fitting phase is the crucial one, as it implicitly determines the quality of the subsequent clustering results. However, it represents a very hard task; indeed, the problem of fitting a mixture model is always difficult, but it is even more severe in OID applications. This is due to the huge class-skew inherent in OID problems, where the very few (and unidentified) distance measures stemming from matches risk to be completely overwhelmed by the bulk of those stemming from non-matches. To overcome this difficulty MAERLIN exploits an original fitting technique inspired by perturbation theory (see, e.g., [53]) and designed to take advantage from both PK1 and PK2. The technique is coded as a two-step algorithm, with the M class mixing weight playing the role of the perturbative expansion parameter. The first-step concentrates on the U component mixture parameters and is specifically aimed at "factorizing" the leading contribution arising from non matches. The second-step strives to increase the Likelihood achieved in the previous step by using the remaining mixture parameters in a smart way; that is, M density parameters are tuned in such a way as to better fit the behaviour of the distance distribution exactly in those regions where, thanks to PK1, values stemming from matches are more likely to be found.

In the clustering phase MAERLIN searches an optimal classification rule such that each pair can be assigned, based on its observed distance value, either to the M or to the U class. The system can minimize either the probability of classification error (maximum likelihood objective) or, alternatively, the expected classification cost (minimum cost objective), while satisfying arbitrary matching constraints among the two sets of objects to be matched (1:1, 1:n, n:1 or n:m). If no constraints are imposed (i.e. for n:m matching), the applied classification rules depend in a quite straightforward way on posterior estimates of class membership probabilities and reflect classical decision theory results (see, e.g., [187]). For instance, the maximum likelihood objective leads to the well known Maximum a Posteriori (MAP) rule, see Figure 1.3. When, on the contrary, matching constraints are imposed, MAERLIN faces directly the full-complexity constrained optimization problem by means of a purposefully designed evolutionary algorithm [689].

1.3.1.2 Need for time-aware techniques

Let's consider the second aspect related to OID and time-dependance, i.e. the possible availability of a timestamp for Web data. The OID matching process does need to be aware of this specific kind of information, and indeed there are some preliminary works that actually take explicitly into account the temporal information. As an example, in [496], an approach that leverages temporal information with linkage is presented. The approach takes into account cases in which as time elapses, values of a particular entity may evolve; for example, a researcher may change affiliation or email. On the other hand, different objects are more likely to share the same value(s) with a long time gap. Thus the concept of *decay* is defined, with which the penalty for value disagreement is reduced and, at the same time, the reward for value agreement over a long period is reduced as well. Moreover, temporal clustering algorithms are proposed that explicitly consider time order of records in order to improve linkage results.

1.3.2 Object Identification and Quality

When considering OID of Web data, quality becomes a fundamental issue: the complexity of the process is the greater the more the quality of data is poor. Assessing the quality of Web data is a current research activity, and is of course highly dependent on the specific Web source. In the following, we give two examples that show that, unfortunately, the overall quality of Web data appears to be dramatically poor.

A first example is related to social media data, such as Twitter data. As reported in [90] Twitter has been used to examine a wide variety of patterns such as mood rhythms, media event engagement, political uprisings, etc. However,

"Twitter does not represent "all people", and it is an error to assume "people" and "Twitter users" are synonymous: they are a very particular sub-set. Neither is the population using Twitter representative of the global population. Nor can we assume that accounts and users are equivalent. Some users have multiple accounts, while some accounts are used by multiple people. Some people never establish an account, and simply access Twitter via the web. Some accounts are "bots" that produce automated content without directly involving a person."

Twitter data are characterized for being highly unstructured, and often not accompanied by metadata. This means that high percentages of these data cannot be simply used by automated processes, as they are "pointless babbles" [89]. To get an effective use of this kind of data it is necessary to investigate methods for automatic generation of the right metadata to describe the data under review.

The second example of Web quality assessment is related to deep Web data. Deep Web indicates that part of the Web that is not directly indexed by standard search engines. A huge amount of information on the Web is sunk on dynamically generated sites, and traditional search engines cannot access this information as those pages do not exist until they are created dynamically as the result of a specific search. Most Web sites are interfaces to databases, including e-commerce sites, flight companies sites, online bibliographies, etc. The deep Web includes all these sites and thus it is estimated that its size is several orders of magnitude larger than the surface Web [58].

In [402], an assessment of the quality of deep Web data from stock (55 sources) and flight (38 sources) domains is presented. The results of the assessment report a bad quality in terms of inconsistency (for 70% of data items more than one value is provided) and of correctness (only 70% correct values are provided by the majority of the sources).

Interestingly, the work [402] provides a specific definition of quality metrics for Web data. Such set of metrics is described in the following and represents one of the first attempts to define a quality assessment framework for Web data.

First, it is evaluated the *redundancy* of data. Specifically: (i) redundancy on objects, i.e. the percentage of sources that provide a particular object, and (ii) redundancy on data items, i.e. the percentage of sources that provide a particular data item.

The further considered dimension is *consistency* of the data, defined according to three measures.

- Number of values. By denoting as $V(d)$ the set of values provided by various sources on d, number of values reports the number of different values provided on d, that is the size of $V(d)$.
- Entropy. By denoting as $S(d)$ the set of sources that provide data on item d, and $S(d;v)$ the set of sources that provide value v on d, the entropy is

  $\sum_{v \in V(d)} \frac{|S(d,V)|}{|S(d)|} \log \frac{|S(d,V)|}{|S(d)|}$.

  Intuitively, the higher the inconsistency, the higher the entropy.
- Deviation. For numerical values, by defining as v_0 the dominant value, i.e. the one with the largest number of providers given by $\text{argmax}_{v \in V(d)} |S(d,V)|$, the deviation from d is:

  $D(d) = \sqrt{\frac{1}{|V(d)|} \sum_{v \in V(d)} (\frac{v - v_0}{v_0})^2}$

Finally, accuracy is evaluated according to two measures, namely:

- Source accuracy: We compute accuracy of S as the percentage of its provided values that are consistent with the given gold standard.
- Accuracy deviation: computed as the standard deviation of the accuracy of a source over a period of time. Given that $\mathcal{T}$ is the set of time points in a period, $A(t)$ is the accuracy of the source at time $t \in \mathcal{T}$, and $\bar{\mathcal{A}}$is the mean accuracy over $\mathcal{T}$, the variety is computed by: $\sqrt{\frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} (A(t) - \bar{\mathcal{A}})^2)}$

The final results of the assessment activity performed according to such measures are quite poor, namely:

- For the stock domain, there is a very high redundancy at the object level, namely each source provides over 90% of the stocks; for the flight domain, object-level redundancy is lower, namely only 36% of the sources cover 90% of the flights. It is observed that there is large redundancy on data items, over various domains: on average each data item has a redundancy of 66% for stock and 32% for flight.
- There is a quite high inconsistency of values on the same data item: for stock and flight the average entropy is .58 and .24, and the average deviation is 13.4 and 13.1 respectively. The inconsistency can vary from attributes to attributes. By choosing dominant values as the true value precision is 0.908 for stock and 0.864 for flight for the two domains respectively.
- Accuracy of the sources can vary a lot: on average the accuracy is about .86 for stock and .80 for flight.

1.4 Quality of Big Data: a Classification of Big Data Sources

The term Big Data (BD) is used for identifying structured or unstructured data sets that are impossible to store and process using common software tools (e.g. relational databases), regardless of the computing power or the physical storage at hand. The size of data, typically spanning dimensions of tera and peta bytes orders of magnitude, is not the only aspect that make data "Big". Indeed, the problem of feasibility in treating data increases when data sets grow continuously over time while a timely processing is necessary for producing business value [550]. According to a classification proposed by UNECE (United Nations Economic Commission for Europe) (see [615]), there are three main types of data sources that can be viewed as Big Data:

1. human-sourced information sources;
2. process-mediated sources;
3. machine-generated sources.

Type 1 sources include a vast amount of data types such as: a. social networks (Facebook, Twitter, LinkedIn, etc.), b. blogs and comments, c. internet searches on search engines (Google, etc.), d. videos loaded in the Internet (You Tube, etc.), e. user-generated maps, f. picture archives (Instagram, Flickr, Picasa, etc.), g. data and contents from mobile phones (text messages, etc.), h. e-mails, and so on.

Type 2 sources can consist of: a. data produced by public bodies and institutions (medical records, etc.), and b. data produced by the private sector (commercial transactions, banking/stock records, e-commerce, credit cards, etc.).

Among type 3 sources, we can distinguish: a. data from fixed sensors (home automation, weather/pollution sensors, traffic sensors/web cameras, scientific sensors, security/surveillance videos/images, etc.), b. data from mobile sensors, i.e. for tracking or analysis purposes (satellite images, GPS, mobile phone location, car devices, etc.), and c. data from computer systems (log files, web logs, etc.).

Big Data is gaining more and more attention both in academic and business contexts. Nowadays, the main unmatched challenges in Big Data management concern the so called 3V:

- *variety*, referring to the heterogeneity of data acquisition, data representation, and semantic interpretation. As to BD representations, we have introduced in the Preface two evolution coordinates for information types, the perceptual coordinate and the linguistic coordinate.
- *volume*, referring to the size of the data. Worldwide information volume is growing at a rate of 60 % annually, and 90 % of data in the today world has been created during the last two years.

- *velocity*, referring to the data provisioning rate and to the time in which it is necessary to act on them. Every minute 400.000 tweets on Tweet are posted, 200 millions of e-mails are sent, 2 millions of Google search queries are submitted [461].

Given that BD involves so many different sources and business domains, a quality characterization of them should be *source-specific* and *domain-specific*.

Source-specificity is very much evident when considering the heterogeneous nature of some sources. For instance, sensor network's data streams can be quality characterized by the fact that data is often missing, and when not missing they are subject to potentially significant noise and calibration effects. In addition, because sensing relies on some form of physical coupling, the potential for faulty data is high. Depending on where a fault occurs in the data reporting, observations might be subject to unacceptable noise levels (for example, due to poor coupling or analog-to-digital conversion) or transmission errors (packet corruption or loss). In Section 1.5.1 we will discuss in detail quality issues in sensor data sources.

Conversely, for social media data, data are highly unstructured, and often not accompanied by metadata. This means that high percentages of these data cannot be simply used by automated processes as they are affected by high percentages of noise. In the other cases, however, dedicated and often expensive activities of semantic extraction must be performed.

Domain-specificity is the other relevant dimension for the specific characterization of quality of Big Data. Depending on the domain, it is necessary to focus on some aspects of Big Data quality rather than others. In Section 1.6 we will see the example of the Official Statistical domain for which the representativeness or selectivity of Big Data sources is a particularly relevant feature. Indeed, statistical production processes do have to seriously take into account such a feature in order to produce reliable estimates.

1.5 Source-specific Quality Issues in Sensor Data

Big Data sources of type 3 include sensors and sensor networks (S&SN). In this section, we first discuss the evolution of the S&SN technology, and the most relevant applications. Then, we consider the most usual fault events and phenomena that affect IQ. We also analyse quality dimensions that are characteristic of this technology, and some techniques proposed for quality assessment and improvement.

1.5.1 Information Quality in Sensors and Sensor Networks

Sensor networks can be defined as large-scale ad hoc networks of homogeneous or heterogeneous, compact, mobile or immobile sensor nodes that are randomly deployed in an area of interest [256]. Different types of data are collected by the sensor nodes, e.g. application specific environmental parameters, meteorological or Global Positioning System. These data can be in different forms, digital and analogue, spatial and temporal, alphanumeric or image, fixed or moving. The measurements taken by the sensor nodes in SN are discrete samples of physical phenomena that are subject to review of their accuracy dependent on location. General causes of errors in sensor data include: a. noise from external sources, b. hardware noise, c. inaccuracies and impressions in sampling methods and derived data, and d. environmental effects. In addition, corruption of functioning can result from e. adverse weather conditions, f. faulty equipment, or g. human error.

[372] observes that the underlying measurement process as well as sensor failures or malfunctions may lead to falsified, wrong, or missing values. To extract complex knowledge, sensor data are merged, transformed, and aggregated by applying traditional data stream queries, complex signal analysis, or numerical operators. During the data stream processing task, the initial sensor-inherent errors may be amplified. Additionally, new errors may be introduced.

For [346], "dirty data" in receptor data manifest themselves in three general forms: a. missed readings, for example, RFID readers often capture only 60-70% of the tags in their vicinity; b. unreliable readings: often, individual sensor readings are imprecise and/or unreliable; c. variance in errors due to the environment.

When data are collected in S&SN, their quality can deeply impact on decisions to be taken, e.g.:

- data may not be readily available for analysis and interpretation;
- problems with the equipment, such as battery voltage, high differences between the temperature of the instrument and the external temperature, and dark current drifts, might be difficult to identify;
- as the complexity of the equipment increases, so does the difficulty to determine the cause of equipment malfunctions.

Besides general descriptions of quality of information in S&SN, in the following we report two proposals of quality dimensions:

- [561], detailing S&SN quality dimensions as subtypes of consistency;
- [428], linking quality of S&SN to the notion of *Quality of Context (QoC)*.

[561] defines several subtypes of consistency, shown in Table 1.4, together with their definitions and an identification whether the dimension refers to individual data or data streams. At a macro level, three types of consistency

are considered, namely: numerical, temporal and frequency consistency: numerical consistency is equivalent to accuracy; temporal consistency is to be meant as a degree of up-to-dateness; frequency consistency focuses on abnormal changes in data provisioning.

[428] observes that diverse sources of context information, ranging from physical and logical sensors to user interfaces and applications on mobile devices, affect the quality of context data. QoC sources are the information about the sources that collect context information, the environments where that context information is collected, and the entities about which the context information is collected. Examples of QoC sources are: source location, measurement time, source state, sensor data accuracy etc..

QoC parameters are derived from QoC sources and are represented in a form that is suitable for use by an application. QoC parameters can be divided into generic and domain specific parameters. Generic QoC parameters are those parameters which are required by most applications, such as up-to-dateness, trustworthiness, completeness, representation consistency, and precision. Domain specific QoC parameters are those parameters that are important for some specific application domains. Table 1.5 summarizes the main concepts introduced in [428]: on one side dimensions with clusters they belong to and their definitions, and, on the other side related QoC sources.

Types of Consistency	**Numerical/ Temporal/ Frequency**	**Individual Data/ Data Streams/ Both**	**Definition**
Numerical	Numerical	Individual Data	Collected data should be accurate
Temporal	Temporal	Individual Data	Data should be delivered to the sink before or by it is expected
Frequency	Frequency	Both	Controls the frequency of dramatic data changes and abnormal readings of data streams
Absolute numerical	Numerical	Both	Sensor reading is out of the normal range, which can be preset by the application
Relative numerical	Numerical	Both	Error between the real field reading and the corresponding data at the sink
Hop	Numerical	Individual Data	Data should keep consistency at each hop
Single path	Numerical and Temporal	Individual Data	Consistency holds when data are transmitted from the source to the sink using a single path
Multiple path	Numerical and Temporal	Individual Data	Consistency holds when data are transmitted from the source to the sink using multiple paths
Strict	Numerical and Temporal	Data Streams	Differs from hope consistency because it is defined on a set of data and requires no data loss
Alpha-loss	Numerical and Temporal	Data Streams	Similar to strict consistency except that alfa-data loss are accepted at the sink
Partial	Numerical and Temporal	Data Streams	Similar to alfa consistency except that temporal consistency is released
Trend	Numerical and Temporal	Data Streams	Similar to partial consistency except that numerical consistency is released
Range frequency	Frequency	Data Streams	The number of abnormal readings exceed a certain number preset by the application
Change frequency	Frequency	Data Streams	Changes of sensor readings exceeds preset threshold

Table 1.4 Various types of consistency as defined in [561]

.

1.5.2 Techniques for Data Cleansing in Sensors and Sensor Networks

A variety of techniques are currently investigated for IQ management in S&SN.

[347] observes that the nature of the errors in receptor data is not easily corrected by traditional data cleaning. Receptor data demands different techniques that address the nature of its errors (i.e., missed and unreliable readings). These data tend to be strongly correlated in both time and space; the readings observed at one time instant are highly indicative of the readings observed at the next time instant, as are readings at nearby devices. To provide a simple and flexible means of programming cleaning tools, [348] proposes to specify cleaning stages using high-level declarative queries over relational data streams; the system then translates the queries into the appropriate low-level operations necessary to produce cleaned results.

As to dimensions and techniques for specific sensor technologies such as RFID, [348] observes that one of the primary factors limiting the widespread adoption of RFID technology is the unreliability of the data streams produced by RFID readers. To face with such an issue, a temporal "smoothing filter" is proposed, namely a sliding window over the reader's data stream that interpolates for lost readings from each tag within the time window. The goal is to reduce or eliminate dropped readings by giving each tag more opportunities to be read within the smoothing window. Unlike conventional techniques, the technique does not expose the smoothing window parameter to the application; instead, it determines the most appropriate window size automatically and continuously adapts it over the lifetime of the system based on observed readings. [516] discusses the issue of dealing with anomalies in RFID reads, where each application specifies the detection and the correction of relevant anomalies using declarative sequence-based rules.

The contributions of [126] concern spatial redundancy (and consequent spatial inconsistency), where an object is detected by multiple readers in its neighborhood, and temporal redundancy (and consequent temporal inconsistency), where an object is detected multiple times by a single reader over time.

Finally, as to IQ and the new frontier of participatory sensing in social networks, [100] observes that mobile devices are increasingly capable of capturing, classifying and transmitting image, acoustic, location and other data, interactively or autonomously. They could act as sensor nodes and location-aware data collection instruments. [100] introduces the concept of *participatory sensing*, which asks everyday mobile devices, such as cellular phones, to form interactive, participatory sensor networks that enable public and professional users to gather, analyze and share local knowledge.

Cluster	Dimension in Cluster	Definition	Sources of QoC used in the evaluation
Accuracy	Up-to-Dateness	Degree of rationalism to us e acontext object for a specific application at a given time	Measurement Time Current Time
Accuracy	Precision	–	–
Completeness	Completeness	Quantity of information that is provided for a specific object	Ratio of number of attributes filled to the total number of attributes
Completeness	Significance	Worth or preciousness of the context information in a specific situation	Critical value
Redundancy	Conciseness	–	–
Consistency	Representation Consistency	–	–
Trustworthiness	Trustworthiness	Belief that we have in the correct information in a given context object	Source location Information entity location Sensor data accuracy

Table 1.5 Clusters, quality of context dimensions, definitions in [428] and related sources of context data

.

1.6 Domain-specific Quality Issues: Official Statistics

In this section we discuss Big Data quality issues in the Official Statistics (OS) domain.

The main purpose of official statistics is well-defined by Principle 1 of the Fundamental Principles of Official Statistics, as provided by the UN Statistics Division [179]:

Official statistics provide an indispensable element in the information system of a democratic society, serving the Government, the economy and the public with data about the economic, demographic, social and environmental situation.

The quality of data resulting from OS production by National Statistical Institutes is therefore a primary issue. National Statistical Institutes started investigating the roles that Big Data can have in Official Statistics either for use on its own, or in combination with more traditional data sources such as sample surveys and administrative registers [268]. Recently, the Scheveningen memorandum [436], which has the role of providing strategic guidelines to European national offices, clearly stated that, given the opportunities that Big Data offer to OS, National Statistical Institutes are encouraged to undertake initiatives to examine the potential of Big Data sources in that regard. In the following, we first define the concept of quality of Big Data fo OS (Section 1.6.1), then, in Section 1.6.2 we describe a case study showing examples of quality issues that can emerge when conducting a Big Data project in the OS domain.

1.6.1 On the Quality of Big Data for Official Statistics

There are a number of issues that are specific of the OS domain, mainly:

- Selectivity and representativeness: populations covered by Big Data sources are not typically the target populations of OS and are often not explicitly defined. Moreover, given that the Big Data generating mechanisms are not under OS control, data deriving from Big Data sources can be selective, i.e. not representative of the target population. Dealing with these issues is not easy, especially because it is not always feasible to assess the relationships between the covered population and the target population on one side, and to estimate the bias to control, on the other side.

- Data processing: This issue is concerned with three different aspects that are very important for dealing with Big Data in OS, namely: (i) data preparation, (ii) data filtering, (iii) data reconciliation. With respect to (i), big sources are typically event-based rather than unit-based, as it traditionally happens for OS survey data (or for administrative data). Hence a first preparation step is needed in order to deal with such new types of data. With respect to (ii) Big Data are often affected by "noise" with respect to the analysis purpose, that must be filtered.
  On one side, this noise is related to the fact the data generation process is not under a direct control of the statistician that cannot apply a "design" to the data collection phase. On the other side the noise can be related to particular nature of some sources, like unstructured information sources (e.g. Twitter data). With respect to (iii), even when some schema or metadata information is present in Big Data sources, such metadata need to be reconciled with metadata driving the statistical production, hence a reconciliation step is needed. As a further observation, due to the great variety of schema information that can derive from Big Data sources (e.g. Internet data), the reconciliation step can be very hard due to the sparsity/incompleteness of Big Data sources schemas.
- Quality of estimates: this issue is related to the major paradigm shift in the analysis activities caused by the usage of Big Data. In particular, data analysis approaches traditionally used within OS may not be directly applied to Big Data analysis. Methodologies that proceed by exploratory analysis, like those based on data mining and machine learning, could be, instead, more appropriately applied. However, they are new for OS: though they are currently successfully applied in specific domains (e.g. customer profiling), their usage in the OS domain has still to be properly investigated.
- Integration with traditional data sources: this issue is related to the usage of Big Data sources integrated with survey-based data or administrative data sources. However, several problems have been identified: (i) linking Big Data is hard because of privacy issues that prevent Big Data vendors to release data that are identifiable; (ii) integration task requires to have a precise and explicit structural metadata representation (schema information) that is often not available for Big Data; (iii) even when schema information is available, it will need to be reconciled with traditional sources schemas.

In the following we describe a case study showing a concrete usage of Big Data for official statistics by focusing on quality-related issues.

1.6.2 A Case Study

Among the different possible types of Big Data sources, Internet data are surely among the most at hand and promising; Internet As a Data source (IaD) has been more and more emerging as a paradigm that concretely allows to complement or substitute traditional statistical sources that, for official statistics, are either resulting from surveys questionnaires or from administrative sources.

In this section, we describe an experimental project conducted by Istat, the Italian National Institute of Statistics, adopting IaD for collecting data. The project has been carried out within the Istat sampling survey on "ICT in enterprises" that aims at producing information on the use of ICT and in particular on the use of Internet by Italian enterprises for various purposes (e-commerce, e-recruitment, advertisement, e-tendering, e-procurement, e-government). To do so, data are collected by means of the traditional instrument of the questionnaire.

Istat started to explore the possibility to use Web scraping techniques, associated in the estimation phase with text and data mining algorithms, in order to substitute traditional instruments of data collection and estimation, or to combine them in an integrated approach. Hence, in the project, the 8600 Web sites, indicated by the 19000 respondent enterprises, have been scraped; acquired texts were processed in order to estimate information which is currently collected via questionnaires.

As described in [542], the overall process consisted of the following phases:

- *web scraping*: aimed at transforming the (unstructured) information in each web site into indexed documents that can be stored and analysed;
- *terms extraction and normalization*: targeted to identify those terms that could provide information on the Internet usage by enterprises;
- *inference activity*: aimed at estimating some classification models in order to come up with estimated answers to questionnaires, derived from enterprises' Web sites.

The inference activity of the process is particularly relevant for the quality aspects and is described in the following as reported in [35]. The input to the inferential activity was a document/term matrix, where each row represents a website, each column is referred to an influent word, and the intersection indicates the presence or the absence of the word in the website.

In order to choose the best instruments useful to build the inference system, in this exploratory phase several tools were tested namely:

- data mining learners, applicable to this text mining problem: *Classification Trees*, *ensemble learners* (*Random Forest, Adaptive Boosting, Bootstrap Aggregating*), *Neural Networks, Maximum Entropy, Support Vector Machines, Latent Dirichlet Allocation* ([339]);
- a text mining learner: *Naïve Bayes* ([387]);
- the approach followed in the *Content Analysis* ([317]).

As usual, available data have been partitioned into a training set and a test set: each model, fitted using the training set, has been applied to the test set in order to evaluate its performance, by comparing observed and predicted values for the target variables, both at individual and aggregate level. In general, the proportion between the two sets was determined in 75/25, but a sensitivity analysis has been performed for Naïve Bayes and content analysis defining nine different levels for the training set (from 10% to 90%). Experiments have been carried out considering the four different subsets of words defined accordingly to their chi-square, and the most favorable in terms of performance has been retained. Performance has been measured by considering the following indicators: (i) *precision* (number of correctly classified cases on the total number of cases), (ii) *sensitivity* (rate of correctly classified positive cases), (iii) *specificity* (rate of correctly classified negative cases). Besides, (iv) the *proportion of predicted positive cases* was introduced, as it corresponds to the final estimates needed and whose accuracy was important to maximize.

From such a comparative analysis, the best method among those considered resulted to be Naïve Bayes. This method was applied in order to estimate other suitable variables in the questionnaire, obtaining the results reported in table 1.6.

Table 1.6 Results of the application of Naïve Bayes to the complete set of questions related to Web sales.

Question	Precision	Sensitivity	Specificity	Proportion Web sales = Yes (observed)	Proportion Web sales = Yes (predicted)
Web sales functionality	0.78	0.50	0.86	0.21	0.21
Orders tracking	0.82	0.49	0.85	0.18	0.11
Description and price list of goods	0.62	0.44	0.79	0.48	0.32
Personalised content for regular visitors	0.74	0.41	0.781	0.09	0.23
Possibility to customise online goods	0.86	0.53	0.87	0.05	0.14
Privacy policy statement	0.59	0.57	0.64	0.68	0.51
Online job application	0.69	0.521	0.78	0.35	0.33

The final obtained results can be considered satisfiable. Interestingly, in some cases it was possible to verify by manual inspection that some enterprises answering *no* to the web sales, do provide instead web sales, i.e. the answer was probably due to a misunderstanding of the question. In these cases, the automatic approach even outperforms the traditional one with respect to quality of the answers. Anyway, an issue related to the quality of estimates is there: the adoption of machine learning approaches, not typically used in OS, poses the issue to verify the reliability of the results. The study described in this section is a step towards such kind of verification. Once this alternative approach will be proved to offer a quality of obtainable estimates higher than that of the traditional approach, the new process could become an important part of the survey on "ICT in enterprises". It will also be possible to consider not only an improvement of the accuracy of already available estimates, but also to produce new estimates related to additional information currently not covered by the survey. Finally, in order to detect erroneous values in the survey data, predicted values could be used in the editing phase of the current production process.

1.7 Summary

In order to exploit the enormous range of opportunities deriving from using Web data, it is really important to have a quality characterization of them. Trustworthiness and provenance, as discussed in this chapter, have a prominent role in such a characterization. Web data often need to be integrated with more traditional data sources that are typically used in business processes. To the scope, activities like matching Web data objects assume a particularly important role.

The discussion on issues and techniques related to Web information quality as performed in this chapter has been complemented with consideration on Big Data quality. This is a huge and hot issue: on one side, at this stage, it seems really mandatory the use of Big Data as a source of information, on the other side it is necessary to characterize the quality of Big Data properly, in order to make a correct use of it. As shown by the examples on the quality of sensor data and on quality issues in Official Statistics, however, the way to well-defined methods and approaches has been just taken and is still a bit long to go.

References

[1] A Mittal AKM, Bovik AC (2012) No-reference image quality assessment in the spatial domain. IEEE Transactions on Image Processing 21(12):4695–4708

[2] Aalst vdW, Hofstede tA (2005) YAWL: Yet Another Workflow Language. Information Systems 30(4):245–275

[3] AAVV (2004) Information and data quality in the NHS. Tech. rep., UK Audit Commission, London, UK, URL http://archive.audit-commission.gov.uk/auditcommission/SiteCollectionDocuments/AuditCommissionReports/NationalStudies/20040330dataquality.pdf

[4] Abdelhak M, Grostick S, Hanken MA (eds) (2012) Health information: management of a strategic resource, 4th edn. Elsevier Saunders, St. Louis, MO

[5] Abiteboul S, Buneman P, Suciu D (2000) Data on the Web: From Relations to Semistructured Data and XML. Morgan Kaufmann Publishers

[6] Abowd JM, Vilhuber L (2005) The sensitivity of economic statistics to coding errors in personal identifiers. Journal of Business & Economic Statistics 23(2)

[7] Adams S, Berg M (2004) The nature of the net: constructing reliability of health information on the web. Information Technology & People 17(2):150–170, DOI 10.1108/09593840410542484, URL http://www.emeraldinsight.com/10.1108/09593840410542484

[8] Adams SA (2010) Revisiting the online health information reliability debate in the wake of web 2.0: an inter-disciplinary literature and website review. International Journal of Medical Informatics 79(6):391–400

[9] Agnoloni T, Francesconi E (2011) Modelling semantic profiles in legislative documents for enhanced norm accessibility. In: ICAIL, pp 111–115

[10] Agrawal R, Gupta A, Sarawagi S (Birmingham U.K, April 7-11, 1997) Modeling multidimensional databases. In: Gray A, Larson P (eds) Proceedings of the 16th International Conference on Data Engineering (ICDE 2000), IEEE Computer Society, pp 232–243

[11] Ahituv N (1980) A systematic approach toward assessing the value of an information system. MIS Quarterly 4(4)

[12] Ahituv N (1987) Assessing the value of information: Problems and approaches. Tel Aviv University, Faculty of Management, The Leon Recanati Graduate School of Business Administration

[13] Ahituv N, Igbaria M, Sella A (1998) The effects of time pressure and completeness of information on decision making. Journal of Management Information Systems 15(2):153–172

[14] Aizawa A, Oyama K (2005) A fast linkage detection scheme for multi-source information integration. In: Web Information Retrieval and Integration, 2005. WIRI'05. Proceedings. International Workshop on Challenges in, IEEE, pp 30–39

[15] Al-Lawati A, Lee D, McDaniel P (2005) Blocking-aware private record linkage. In: Proceedings of the 2nd international workshop on Information quality in information systems, ACM, pp 59–68

[16] Altowim Y, Kalashnikov DV, Mehrotra S (2014) Progressive approach to relational entity resolution. Proceedings of the VLDB Endowment 7(11)

[17] Altwaijry H, Kalashnikov DV, Mehrotra S (2013) Query-driven approach to entity resolution. Proceedings of the VLDB Endowment 6(14):1846–1857

[18] Aluisio S, Specia L, Gasperin C, Scarton C (2010) Readability assessment for text simplification. In: Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics, pp 1–9

[19] Amann B, Constantin C, Caron C, Giroux P (2013) Weblab prov: computing fine-grained provenance links for xml artifacts. In: EDBT/ICDT Workshops, pp 298–306

[20] Amat G, Laboisse B (18th January 2005, Paris, France) B.d.q.s. une gestion opérationnelle de la qualité de données. In: First Data and Knowledge Quality Workshop, In conjunction with ECG

[21] Amsterdam AU (2001) The role of verification in improving the quality of legal decision-making. In: Legal Knowledge and Information Systems: JURIX 2001: the Fourteenth Annual Conference, IOS Press, vol 70

[22] Anand MK, Bowers S, Ludscher B (2010) Techniques for efficiently querying scientific workflow provenance graphs. In: In International Conference on Extending Database Technology (EDBT), pp 287–298

[23] Ananthakrishna R, Chaudhuri C, Ganti V (Hong Kong, China, 2002) Eliminating Fuzzy Duplicates in Data Warehouses. In: Proceedings of VLDB 2002

[24] Arasu A, Chaudhuri S, Kaushik R (2008) Transformation-based framework for record matching. In: Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on, IEEE, pp 40–49

[25] Arasu A, Chaudhuri S, Kaushik R (2009) Learning string transformations from examples. Proceedings of the VLDB Endowment 2(1):514–525

[26] Arasu A, Götz M, Kaushik R (2010) On active learning of record matching packages. In: Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, ACM, pp 783–794

[27] Arenas M, Bertossi LE, Chomicki J (1999) Consistent Query Answers in Inconsistent Databases. In: Proc. PODS'99

[28] Arts DG, De Keizer NF, Scheffer GJ (2002) Defining and improving data quality in medical registries: a literature review, case study, and generic framework. Journal of the American Medical Informatics Association 9(6):600–611

[29] Asher J, Fienberg SE, Stuart E, Zaslavsky A (2003) Inferences for finite populations using multiple data sources with different reference times. In: Proceedings of Statistics Canada Symposium 2002: Modelling Survey Data For Social and Economic Research. Statistics Canada, vol 385

[30] Atzeni P, de Antonellis V (1993) Relational Database Theory. The Benjamin /Cummings Publishing Company, Inc.

[31] Ballou D, Wang R, Pazer H, Tayi G (1998) Modeling information manufacturing systems to determine information product quality. Management Science 44(4)

[32] Ballou DP, Pazer HL (1985) Modeling data and process quality in multi-input, multi-output information systems. Management science 31(2):150–162

[33] Ballou DP, Pazer HL (2003) Modeling completeness versus consistency tradeoffs in information decision contexts. IEEE Transactions on Knowledge Data Engineering 15(1):240–243

[34] Ballou DP, Tayi GK (1999) Enhancing data quality in data warehouse environments. Communications of the ACM 42(1):73–78

[35] Barcaroli G, Nurra A, Scarno M, Summa D (2014) Use of web scraping and text mining techniques in the istat survey on information and communication technology in enterprises. In: Proceedings of Quality Conference 2014 (Q2014), Wien, Austria, 2014

[36] Bartleson C (1982) The combined influence of sharpness and graininess on the quality of color prints. Journal Photogr Sci pp 33–38

[37] Batini C, Scannapieco M (2006) Data Quality: Concepts, Methodologies and Techniques (Data-Centric Systems and Applications). Springer-Verlag New York, Inc.

[38] Batini C, Ceri S, Navathe S (eds) (1992) Conceptual Data Base Design: An Entity Relationship Approach. Benjamin and Cummings

[39] Batini C, Barone D, Cabitza F, Ciocca G, Marini F, Pasi G, Schettini R (2008) Toward a unified model for information quality. In: Proceedings of the International Workshop on Quality in Databases and Man-

agement of Uncertain Data, Auckland, New Zealand, August 2008, pp 113–122

[40] Batini C, Cabitza F, Cappiello C, Francalanci C (2008) A comprehensive data quality methodology for web and structured data. To appear in International Journal of Innovative Computing and Applications

[41] Batini C, Cappiello C, Francalanci C, Maurino A (2009) Methodologies for data quality assessment and improvement. ACM Computing Surveys (CSUR) 41(3):16

[42] Batini C, Grega S, Maurino A (2010) Optimal enterprise data architecture using publish and subscribe. In: Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing, ACM, pp 541–547

[43] Batini C, Barone D, Cabitza F, Grega S (2011) A data quality methodology for heterogeneous data. International Journal of Database Management Systems 3(1)

[44] Batini C, Cappiello C, Francalanci C, Maurino A, Viscusi G (2011) A capacity and value based model for data architectures adopting integration technologies. In: A Renaissance of Information Technology for Sustainability and Global Competitiveness. 17th Americas Conference on Information Systems, AMCIS 2011, Detroit, Michigan, USA, August 4-8 2011

[45] Batini C, Palmonari M, Viscusi G (2012) The many faces of information and their impact on information quality. In: Proc. 17th International Conference on Information Quality (IQ 2012), pp 212–228

[46] Batini C, Castelli M, Comerio M, Viscusi G (2014) Value of integration in database and service domains. In: Service-Oriented Computing and Applications (SOCA), 2014 IEEE 7th International Conference on, IEEE, pp 161–168

[47] Batini C, Nardelli E, Tamassia R (April 1986) A Layout Algorithm for Data Flow Diagrams. IEEE Transactions on Software Engineering

[48] Bauer F, Kaltenböck M (2011) Linked open data: The essentials. Edition mono/monochrom, Vienna

[49] Beckett D (2004) RDF/XML Syntax Specification (Revised). Tech. rep., World Wide Web Consortium, `http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210/`

[50] Beeri C, Kanza Y, Safra E, Sagiv Y (2004) Object fusion in geographic information systems. In: Proceedings of the Thirtieth international conference on Very large data bases-Volume 30, VLDB Endowment, pp 816–827

[51] Beeri C, Doytsher Y, Kanza Y, Safra E, Sagiv Y (2005) Finding corresponding objects when integrating several geo-spatial datasets. In: Proceedings of the 13th annual ACM international workshop on Geographic information systems, ACM, pp 87–96

[52] Belin TR, Rubin DB (1995) A Method for Calibrating False Matches Rates in Record Linkage. Journal of American Statistical Association 90:694–707
[53] Bender C, Orszag S (1999) Advanced Mathematical Methods for Scientists and Engineers: Asymptotic methods and perturbation theory. Springer, N. Y.
[54] Benjelloun O, Garcia-Molina H, Menestrina D, Su Q, Whang SE, Widom J (2009) Swoosh: a generic approach to entity resolution. The VLDB Journal, The International Journal on Very Large Data Bases 18(1):255–276
[55] Benson T (2010) Principles of health interoperability HL7 and SNOMED. Springer
[56] Berg M (1999) Accumulating and coordinating: Occasions for information technologies in medical work. Computer Supported Cooperative Work, The Journal of Collaborative Computing 8(4):373–401
[57] Berg M, Toussaint P (2003) The mantra of modeling and the forgotten powers of paper: a sociotechnical view on the development of process-oriented ICT in health care. International journal of medical informatics 69(2):223–234
[58] Bergman MK (2001) The Deep Web: Surfacing Hidden Value. The Journal of Electronic Publishing
[59] Berjawi B (2013) Introduction to the Integration of Location-Based Services of Several Providers
[60] Berndt DJ, Fisher JW, Hevner AR, Studnicki J (2001) Healthcare data warehousing and quality assurance. Computer 34(12):56–65
[61] Berners-Lee T (2006) Design issues: Linked data
[62] Berti-Équille L (2004) Quality-Adaptive Query Processing over Distributed Sources. In: Proc. 9th Internation Conference on Information Quality (IQ 2004)
[63] Berti-Équille L (Yokohama, Japan, 2001) Integration of Biological Data and Quality-driven Source Negotiation. In: Proc. ER 2001
[64] Berti-Equille L, Batini C, Srivastava D (eds) (2005) Exploiting relationships for object consolidation, ACM
[65] Bertolazzi P, Santis LD, Scannapieco M (2003) Automatic record matching in cooperative information systems. In: Proceedings of the ICDT'03 International Workshop on Data Quality in Cooperative Information Systems (DQCIS'03), Siena, Italy
[66] Bertoletti,M and Missier,P and Scannapieco, M and Aimetti, P and C Batini (2005. Shorter version also in ICIQ 2002.) Improving Government-to-Business Relationships through Data Reconciliation and Process Re-engineering. In: Wang R (ed) Information Quality - Advances in Management Information Systems-Information Quality Monograph (AMIS-IQ) Monograph, Sharpe, M.E.
[67] Bhattacharya I, Getoor L (2004) Deduplication and group detection using links. In: KDD workshop on link analysis and group detection

[68] Bhattacharya I, Getoor L (2004) Iterative record linkage for cleaning and integration. In: Proceedings of the 9th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery, ACM, pp 11–18

[69] Bhattacharya I, Getoor L (2007) Collective entity resolution in relational data. ACM Transactions on Knowledge Discovery from Data (TKDD) 1(1):5

[70] Bhattacharya I, Getoor L, Licamele L (2006) Query-time entity resolution. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp 529–534

[71] Bhattacharya S, Sukthankar R, Shah M (2011) A holistic approach to aesthetic enhancement of photographs. ACM Trans Multimedia Comput Commun Appl 7S:21:1–21:21

[72] Biagioli C, Francesconi E, Spinosa P, Taddei M (2003) The nir project: Standards and tools for legislative drafting and legal document web publication. In: Proceedings of ICAIL workshop on e-government: modelling norms and concepts as key issues, pp 69–78

[73] Biagioli C, Cappelli A, Francesconi E, Turchi F (2007) Law making environment: perspectives. In: Proceedings of the V Legislative XML Workshop, pp 267–281

[74] Bianco S, Ciocca G, Marini F, Schettini R (2009) Image quality assessment by preprocessing and full reference model combination. In: Image Quality and System Performance VI, SPIE, vol 7242, p 72420O

[75] Bibliographic Center for Research CDP Digital Imaging Best Practices Working Group (2008) Digital Imaging Best Practices, Version 2.0. Bibliographic Center for Research, URL `http://books.google.it/books?id=vjeEXwAACAAJ`

[76] Bilenko M, Mooney RJ (2003) Adaptive duplicate detection using learnable string similarity measures. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp 39–48

[77] Bilenko M, Kamath B, Mooney RJ (2006) Adaptive blocking: Learning to scale up record linkage. In: Data Mining, 2006. ICDM'06. Sixth International Conference on, IEEE, pp 87–96

[78] Bitton D, DeWitt D (1983) Duplicate record elimination in large data files. ACM Transactions on Databases Systems 8(2)

[79] Bizer C (2007) Quality-driven information filtering in the context of web-based information systems. PhD thesis, Freie Universität Berlin

[80] Black AD, Car J, Pagliari C, Anandan C, Cresswell K, Bokun T, McKinstry B, Procter R, Majeed A, Sheikh A (2011) The impact of eHealth on the quality and safety of health care: a systematic overview. PLoS medicine 8(1):e1000,387

[81] Blakely T, Salmond C (2002) Probabilistic record linkage and a method to calculate the positive predictive value. International Journal of Epidemiology 31(6):1246–1252

[82] Bleiholder J, Naumann F (2008) Data fusion. ACM Computing Surveys
[83] Boag A, Chamberlin D, Fernandez MF, Florescu D, Robie J, Simèon J (2003) XQuery 1.0: An XML Query Language. http:///www.w3.org/TR/xquery
[84] Böhm C, Naumann F, Abedjan Z, Fenz D, Grütze T, Hefenbrock D, Pohl M, Sonnabend D (2010) Profiling linked open data with prolod. In: ICDE Workshops, IEEE, pp 175–178
[85] Bonatti PA, Hogan A, Polleres A, Sauro L (2011) Robust and scalable linked data reasoning incorporating provenance and trust annotations. Journal of Web Semantics 9(2):165 – 201
[86] Bouzeghoub M, Peralta V (Paris, France, June 18th 2004) A framework for analysis of data freshness. In: Proceedings of the International Workshop on Information Quality in Information Systems
[87] Bovee M, Srivastava RP, Mak BR (2001) A Conceptual Framework and Belief-Function Approach to Assessing Overall Information Quality. In: Proc. 6th International Conference on Information Quality (IQ 2001)
[88] Bowers S, McPhillips T, Ludscher B (2012) Declarative rules for inferring fine-grained data provenance from scientific workflow execution traces. In: Intl. Provenance and Annotation Workshop (IPAW), pp 1–15
[89] Boyd D (2009) Twitter: pointless babble or peripheral awareness + social grooming? Tech. rep., Apophenia Inc., URL http://www.zephoria.org/thoughts/archives/2009/08/16/twitterpointle.html
[90] Boyd D, Crawford K (2012) Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information, Communication, & Society 15(5)
[91] Bradley EH, Herrin J, Mattera JA, Holmboe ES, Wang Y, Frederick P, Roumanis SA, Radford MJ, Krumholz HM (2005) Quality improvement efforts and hospital performance: rates of beta-blocker prescription after acute myocardial infarction. Medical care 43(3):282–292
[92] Brandao T, Queluz MP (2008) No-reference image quality assessment based on dct domain statistics. Signal Processing 88(4):822 – 833
[93] Bravo L, Bertossi LE (2003) Logic Programming for Consistently Querying Data Integration Systems. In: Proc. IJCAI 2003
[94] Brickley D, Guha RV (2004) RDF vocabulary description language 1.0: RDF schema. Tech. rep., W3C, http://www.w3.org/TR/2004/REC-rdf-schema-20040210/
[95] Brizan DG, Tansel AU (2006) A survey of entity resolution and record linkage methodologies. Communications of the IIMA 6(3):41–50
[96] Bruni R, Sassano A (2001) Errors Detection and Correction in Large Scale Data Collecting. In: Proceedings of the 4th International Conference on Advances in Intelligent Data Analysis
[97] Buechi M, Borthwick A, Winkel A, Goldberg A (Boston, MA, USA, 2003) ClueMaker: a language for approximate record matching. In: Pro-

ceedings of the 7th International Conference on Information Quality, (ICIQ 2003)
[98] Buneman P (1997) Semistructured data. In: Proceedings of the 16th ACM Symposium on Principles of Database Systems (PODS 1997), Tucson, Arizona
[99] Buneman P, Khanna S, Tan WC (2001) Why and Where: A Characterization of Data Provenance. In: Proceedings of the 8th International Conference on Database Theory (ICDT)
[100] Burke J, Estrin D, Hansen M, Parker A, Ramanathan N, Reddy S, Srivastava MB (2006) Participatory sensing. In: Proceedings of the Workshop on World-Sensor-Web (WSW) at ACM Conference on Embedded Networked Sensor Systems (SenSys 2006), Boulder, Colorado, USA
[101] Byrd LW, Byrd TA (2012) Developing an instrument for information quality for clinical decision making. In: System Science (HICSS), 2012 45th Hawaii International Conference on, IEEE, pp 2820–2829, DOI 10.1109/HICSS.2012.210, URL `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6149169`
[102] Cabitza F (2013) An information reliability index as a simple consumer-oriented indication of quality of medical web sites. In: Pasi G, Bordogna G, Lakhmi J (eds) Quality Issues in the Management of Web Information, Intelligent Systems Reference Library, vol 50, Springer Berlin Heidelberg, pp 159–177
[103] Cabitza F, Simone C (2012) "Whatever works": Making sense of information quality on information system artifacts. In: Viscusi G, Campagnolo GM, Curzi Y (eds) Phenomenology, Organizational Politics, and IT Design: The Social Study of Information Systems, IGI Global, pp 79–110, URL `10.4018/978-1-4666-0303-5.ch006`
[104] Cabitza F, De Michelis G, Simone C (2014) User-driven prioritization of features for a prospective interpersonal health record: Perceptions from the italian context. Computers in Biology and Medicine DOI 10.1016/j.compbiomed.2014.03.009, URL `http://linkinghub.elsevier.com/retrieve/pii/S0010482514000729`
[105] Caldiera VRBG, Rombach HD (1994) Goal question metric paradigm. Encyclopedia of Software Engineering 1:528–532
[106] Cali A, Calvanese D, De Giacomo G, Lenzerini M (2002) On the role of integrity constraints in data integration. IEEE Data Eng Bull 25(3):39–45
[107] Calì A, Lembo D, Rosati R (2003) On the Decidability and Complexity of Query Answering over Inconsistent and Incomplete Databases. In: Proc. PODS 2003
[108] Calì A, Lembo D, Rosati R (2003) Query Rewriting and Answering under Constraints in Data Integration Systems. In: Proc. IJCAI 2003
[109] Callet P, Autrusseau F (2005) Subjective quality assessment IRCCyN/IOVC database. http://www.irccyn.ec-nantes.fr/ivcdb/

[110] Calvanese D, De Giacomo G, Lenzerini M (1999) Modeling and Querying Semi-Structured Data. Networking and Information Systems Journal 2(2):253–273
[111] Cappiello C, Comuzzi M (2009) A utility-based model to define the optimal data quality level in IT service offerings. In: 17th European Conference on Information Systems, ECIS 2009, Verona, Italy, 2009, pp 1975–1986
[112] Carnec M, Callet PL, Barba D (2008) Objective quality assessment of color images based on a generic perceptual reduced reference. Signal Processing: Image Communication 23(4):239 − 256
[113] Carroll J (2003) Signing rdf graphs. Tech. rep., HPL-2003-142, HP Labs
[114] Carroll JG (2004) The gold standard: The challenge of evidence-based medicine and standardization in health care. Quality Management in Healthcare 13(2):150âĂŞ–151
[115] Chall JS (1995) Readability revisited: The new Dale-Chall readability formula, vol 118. Brookline Books Cambridge, MA
[116] Chan KS, Fowles JB, Weiner JP (2010) Review: electronic health records and the reliability and validity of quality measures: a review of the literature. Medical Care Research and Review 67(5):503–527
[117] Chan SY (2001) The use of graphs as decision aids in relation to information overload and managerial decision quality. Journal of information science 27(6):417–425
[118] Chandler D, Hemami S (2007) A57 image database. `http://foulard.ece.cornell.edu/dmc27/vsnr/vsnr.html`
[119] Chandler DM (2013) Seven challenges in image quality assessment: past, present, and future research. ISRN Signal Processing 2013
[120] Charnes A, Cooper W, Rhodes E (1978) Measuring the efficiency of Decision Making Units. European Journal of operational research 2
[121] Chassin MR, Becher EC (2002) The wrong patient. Annals of Internal Medicine 136(11):826–833
[122] Chaudhuri S, Das Sarma A, Ganti V, Kaushik R (2007) Leveraging aggregate constraints for deduplication. In: Proceedings of the 2007 ACM SIGMOD international conference on Management of data, ACM, pp 437–448
[123] Chen CC, Knoblock CA, Shahabi C, Chiang YY, Thakkar S (2004) Automatically and accurately conflating orthoimagery and street maps. In: Proceedings of the 12th annual ACM international workshop on Geographic information systems, ACM, pp 47–56
[124] Chen CC, Shahabi C, Knoblock CA, Kolahdouzan M (2006) Automatically and efficiently matching road networks with spatial attributes in unknown geometry systems. In: the Proceedings of the Third Workshop on Spatio-Temporal Database Management (co-located with VLDB2006), Seoul, Korea, pp 1–8

[125] Chen CC, Knoblock CA, Shahabi C (2008) Automatically and accurately conflating raster maps with orthoimagery. GeoInformatica 12(3):377–410

[126] Chen H, Ku W, Wang H, Sun M (2010) Leveraging spatio-temporal redundancy for rfid data cleansing. In: Proceedings of SIGMOD 2010, Indianapolis, Indiana, USA

[127] Chen Z, Kalashnikov DV, Mehrotra S (2007) Adaptive graphical approach to entity resolution. In: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries, ACM, pp 204–213

[128] Chen Z, Kalashnikov DV, Mehrotra S (2009) Exploiting context analysis for combining multiple entity resolution systems. In: Proceedings of the 2009 ACM SIGMOD International Conference on Management of data, ACM, pp 207–218

[129] Cheney J, Chiticariu L, Tan W (2007) Provenance in Databases: Why, How, and Where. Foundations and Trends in Databases 1:379–474

[130] Chengalur-Smith IN, Ballou DP, Pazer HL (1999) The impact of data quality information on decision making: an exploratory analysis. Knowledge and Data Engineering, IEEE Transactions on 11(6):853–864

[131] Chengalur-Smith IN, Ballou DP, Pazer HL (1999) The impact of data quality information on decision making: An exploratory analysis. IEEE Transactions on Knowledge and Data Engineering 11(6):853–864, DOI http://dx.doi.org/10.1109/69.824597

[132] Chikkerur S, Sundaram V, Reisslein M, Karam L (2011) Objective video quality assessment methods: A classification, review, and performance comparison. Broadcasting, IEEE Transactions on 57(2):165 –182

[133] Chirigati F, Freire J (2012) Towards Integrating Workflow and Database Provenance. In: In 4th International Provenance and Annotation Workshop, IPAW 2012

[134] Chiticariu L, Tan W, Vijayvargiya G (2004) An annnotation management system for relational databases. In: Proceedings of the 30th Very Large Databases Conference (VLDB)

[135] Cho J, Garcia-Molina H (2003) Estimating frequency of change. ACM Trans Internet Technol 3(3):256–290, DOI 10.1145/857166.857170, URL `http://doi.acm.org/10.1145/857166.857170`

[136] Choquet R, Qouiyd S, Ouagne D, Pasche E, Daniel C, Boussaid O, Jaulent MC (2010) The information quality triangle: a methodology to assess clinical information quality. Stud Health Technol Inform 160(Pt 1):699–703

[137] Christen P (2006) A comparison of personal name matching: Techniques and practical issues. In: Data Mining Workshops, 2006. ICDM Workshops 2006. Sixth IEEE International Conference on, IEEE, pp 290–294

[138] Christen P (2007) A two-step classification approach to unsupervised record linkage. In: Proceedings of the sixth Australasian conference on

Data mining and analytics-Volume 70, Australian Computer Society, Inc., pp 111–119

[139] Christen P (2008) Automatic record linkage using seeded nearest neighbour and support vector machine classification. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp 151–159

[140] Christen P (2012) A survey of indexing techniques for scalable record linkage and deduplication. Knowledge and Data Engineering, IEEE Transactions on 24(9):1537–1555

[141] Christen P, Goiser K (2007) Quality and complexity measures for data linkage and deduplication. In: Quality Measures in Data Mining, Springer, pp 127–151

[142] Christen P, Pudjijono A (2009) Accurate synthetic generation of realistic personal information. In: Advances in Knowledge Discovery and Data Mining, Springer, pp 507–514

[143] Christen P, et al (2007) Towards parameter-free blocking for scalable record linkage. Department of Computer Science, Faculty of Engineering and Information Technology, Australian National University

[144] Ciancio A, da Costa A, da Silva E, Said A, Samadani R, Obrador P (2009) Objective no-reference image blur metric based on local phase coherence. Electronics Letters 45(23):1162 –1163

[145] Codd EF (1970) A relational model of data for large shared data banks. Communications of the ACM 13(6):377–387

[146] Cohen WW, Richman J (2002) Learning to match and cluster large high-dimensional data sets for data integration. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp 475–480

[147] Coiera E (2003) Guide to health informatics. CRC Press

[148] Collins SA, Fred M, Wilcox L, Vawdrey DK (2012) Workarounds used by nurses to overcome design constraints of electronic health records. In: NI 2012: Proceedings of the 11th International Congress on Nursing Informatics, American Medical Informatics Association, vol 2012

[149] Consiglio Regionale della Toscana (2003) Indice di qualita': Percorso e metodologia, in italian

[150] Corchs S, Gasparini F, Marini F, Schettini R (2011) Image quality: a tool for no-reference assessment methods. Image Quality and System Performance VIII 7867(1):786712

[151] Corchs S, Gasparini F, Marini F, Schettini R (2012) A sharpness measure on automatically selected edge segments. Image Quality and System Performance IX 8293(1):82930A

[152] Corchs S, Gasparini F, Schettini R (2014) No reference image quality classification for jpeg-distorted images. Digital Signal Processing 30:86–100

[153] Corchs S, Gasparini F, Schettini R (2014) Noisy images-jpeg compressed: subjective and objective image quality evaluation. In:

IS&T/SPIE Electronic Imaging, International Society for Optics and Photonics

[154] Corner BR, Narayanan RM, Reichenbach SE (2003) Noise estimation in remote sensing imagery using data masking. International Journal of Remote Sensing 24(4):689 – 702

[155] Correndo G, Salvadores M, Millard I, Shadbolt N (2010) Linked timelines: Temporal representation and management in linked data. In: [301], URL `http://ceur-ws.org/Vol-665/CorrendoEtAl_COLD2010.pdf`

[156] Cottrell C (2000) Medicare data study spotlights coding errors. Journal of AHIMA/American Health Information Management Association 71(8):58

[157] Crosby P (1979) Quality is free. McGraw-Hill

[158] Crossley SA, Greenfield J, McNamara DS (2008) Assessing text readability using cognitively based indices. Tesol Quarterly 42(3):475–493

[159] Cui Y, Widom J, Wiener JL (2000) Tracing the Lineage of View Data in a Warehousing Environment. ACM Transactions on Database Systems 25(2):179–227

[160] Culotta A, McCallum A (2005) Joint deduplication of multiple record types in relational data. In: Proceedings of the 14th ACM international conference on Information and knowledge management, ACM, pp 257–258

[161] D Jayaraman AM A Mittal, Bovik A (2012) Objective quality assessment of multiply distorted images. In: Proc. of the Asilomar Conference on Signals, Systems and Computers

[162] Damerau FJ (1964) A technique for computer detection and correction of spelling errors. Communications of the ACM 7(3):171–176

[163] Dasu T, Johnson T (2003) Exploratory Data Mining and Data cleaning. J. Wiley Series in Probability and Statistics

[164] Data Warehousing Institute (2005) Data Quality and the Bottom Line: Achieving Business Success through a Commitment to High Quality Data. `http://www.dw-institute.com/`

[165] Datta R, Joshi D, Li J, Wang JZ (2006) Studying aesthetics in photographic images using a computational approach. In: In Proc. ECCV, pp 7–13

[166] Davis GB, Olson MH (1984) Management information systems: conceptual foundations, structure, and development (2nd ed.). McGraw-Hill, Inc., New York, USA

[167] Davis NA (2014) Health information technology, 3rd edn. Elsvier/Saunders

[168] Davis R, Strobe H, Szolovits P (1993) What is knowledge representation. AI Magazine 14(1):17–33

[169] Dayal U (1985) Query processing in a multidatabase system. In: Query Processing in Database Systems, Springer, pp 81–108

[170] De Amicis F, Batini C (2004) A Methodology for Data Quality Assessment on Financial Data. Studies in Communication Sciences

[171] De Giacomo G, Lembo D, Lenzerini M, Rosati R (2004) Tackling Inconsistencies in Data Integration through Source Preferences. In: Proc. IQIS 2004 (SIGMOD Workshop)

[172] De Michelis G, Dubois E, Jarke M, Matthes F, Mylopoulos J, Papazoglou MP, , Schmidt J, Woo C, Yu E (1997) Cooperative Information Systems: A Manifesto. In: Papazoglou M, Schlageter G (eds) Cooperative Information Systems: Trends & Directions, Accademic-Press

[173] De Vries T, Ke H, Chawla S, Christen P (2009) Robust record linkage blocking using suffix arrays. In: Proceedings of the 18th ACM conference on Information and knowledge management, ACM, pp 305–314

[174] Dejaeger K, Hamers B, Poelmans J, Baesens B (2010) A novel approach to the evaluation and improvement of data quality in the financial sector. In: Proceedings of the 15th International Conference on Information Quality

[175] Delic KA, Dayal U (2002) The rise of the intelligent enterprise. Ubiquity 2002(December):6

[176] Dempster A, Laird N, Rubin D (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of Royal Statistical Society 39:1–38

[177] Demter J, Auer S, Martin M, Lehmann J (2012) LODStats – an extensible framework for high-performance dataset analytics. In: EKAW, Springer, LNCS

[178] Dividino R, Sizov S, Staab S, Schueler B (2009) Querying for provenance, trust, uncertainty and other meta knowledge in RDF. Web Semantics: Science, Services and Agents on the World Wide Web 7:204–219

[179] Division UNS ((accessed February 2015)) `http://unstats.un.org/unsd/methods/statorg/FP-English.htm`

[180] Donabedian A (1980) The definition of quality and approaches to its management. Ann Arbor, MI: Health Administration Press

[181] Dong X, Halevy A, Madhavan J (2005) Reference reconciliation in complex information spaces. In: Proceedings of the 2005 ACM SIGMOD international conference on Management of data, ACM, pp 85–96

[182] Dong X, Halevy AY, Madhavan J (2005) Reference Reconciliation in Complex Information Spaces. In: Proc. SIGMOD 2005

[183] Dong XL, Berti-Equille L, Srivastava D (2009) Truth discovery and copying detection in a dynamic world. PVLDB 2(1):562–573

[184] Dovey S, Meyers D, Phillips R, Green L, Fryer G, Galliher J, Kappus J, Grob P (2002) A preliminary taxonomy of medical errors in family practice. Quality and Safety in Health Care 11(3):233–238

[185] Draisbach U, Naumann F (2009) A comparison and generalization of blocking and windowing algorithms for duplicate detection. In:

Proceedings of the International Workshop on Quality in Databases (QDB), pp 51–56

[186] DuBay WH (2004) The principles of readability. Online Submission

[187] Duda R, Hart P, Stork D (2000) Pattern Classification. John Wiley & Sons

[188] Dunn HL (1946) Record Linkage. American Journal of Public Health 36:1412–1416

[189] Durham E, Xue Y, Kantarcioglu M, Malin B (2012) Quantifying the correctness, computational complexity, and security of privacy-preserving string comparators for record linkage. Information Fusion 13(4):245–259

[190] Dusserre L, Quantin C, Bouzelat H (1994) A one way public key cryptosystem for the linkage of nominal files in epidemiological studies. Medinfo MEDINFO 8:644–647

[191] Dykstra RH, Ash JS, Campbell E, Sittig DF (2009) Persistent paper: The myth of "going paperless". In: AMIA Annu Symp Proc. 2009, pp 158–162

[192] Ebell M (1999) Information at the point of care: answering clinical questions. The Journal of the American Board of Family Practice 12(3):225–235

[193] Eckbert M, Bradley A (1998) Perceptual quality metrics applied to still image compression. Signal Processing 70(3):177–200

[194] Eckerson W (2002) Data Quality and the Bottom Line: Achieving Business Success through a Commitment to High Quality Data. Tech. rep., The Data Warehousing Institute

[195] Elfeky MG, Verykios VS, Elmagarmid AK (2002) Tailor: A record linkage toolbox. In: Data Engineering, 2002. Proceedings. 18th International Conference on, IEEE, pp 17–28

[196] Elfeky MG, Verykios VS, Elmagarmid AK (2002) Tailor: A Record Linkage Toolbox. In: Proc. 18th International Conference on Data Engineering

[197] Elfeky MG, Verykios VS, Elmagarmid AK (2002) Tailor: A record linkage toolbox. In: Data Engineering, 2002. Proceedings. 18th International Conference on, IEEE, pp 17–28

[198] Elhadad N, Sutaria K (2007) Mining a lexicon of technical terms and lay equivalents. In: Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing, Association for Computational Linguistics, pp 49–56

[199] Ell B, Vrandečic D, Simperl E (2011) Labels in the web of data. In: Proceedings of the 10th International Conference on The Semantic Web - Volume Part I, Springer-Verlag, Berlin, Heidelberg, ISWC'11, pp 162–176, URL `http://dl.acm.org/citation.cfm?id=2063016.2063028`

[200] Ellingsen G, Monteiro E (2003) A patchwork planet integration and cooperation in hospitals. Computer Supported Cooperative Work, The Journal of Collaborative Computing 12(1):71–95

[201] Elmagarmid AK, Ipeirotis PG, Verykios VS (2007) Duplicate record detection: A survey. Knowledge and Data Engineering, IEEE Transactions on 19(1):1–16

[202] Elmasri R, Navathe S (1994) Foundamentals of Database Systems (5th ed.). Addison-Wesley Publishing Company

[203] Engeldrum PG (2001) Psychometric scaling:avoiding the pitfalls and hazards. In: IS&T's 2001 PICS Conference Proceedings, pp 101–107

[204] English L (2002) Process Management and Information Quality: How Improving Information Production Processes Improves Information (Product) Quality. In: Proc. 7th International Conference on Information Quality (IQ 2002)

[205] English L (2009) Information quality applied: best practices for improving business information, processes, and systems, 1st edn. Wiley Pub., inc, Indianapolis, IN

[206] English LP (1999) Improving Data Warehouse and Business Information Quality. Wiley & Sons

[207] Eppler M, Helfert M (2004) A classification and analysis of data quality costs. In: ICIS'04: Proceedings of the International Conference on Information Quality, pp 311–325

[208] Eppler MJ (2006) Managing information quality: increasing the value of information in knowledge-intensive products and processes. Springer

[209] Erling O (2012) Virtuoso, a hybrid rdbms/graph column store. IEEE Data Eng Bull 35(1):3–8

[210] European Parliament (2003) Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the Re-use of Public Sector Information. Official Journal of the European Union

[211] European Parliament (2013) Revision of the directive 2003/98/ec of the european parliament and of the council on the re-use of public sector information

[212] EUROSTAT (accessed 2014) `http://ec.europa.eu/eurostat/web/quality/quality-reporting`

[213] EUROSTAT (accessed 2015) `http://epp.eurostat.cec.eu.int/pls/portal/`

[214] Even A, Kaiser M (2009) A framework for economics-driven assessment of data quality decisions. In: AMCIS, p 436

[215] Even A, Shankaranarayanan G (2005) Value-driven data quality assessment. In: IQ

[216] Even A, Shankaranarayanan G (2007) Understanding impartial versus utility-driven quality assessment in large datasets. In: ICIQ, pp 265–279

[217] Even A, Shankaranarayanan G (2007) Utility-driven configuration of data quality in data repositories. International Journal of Information Quality 1(1):22–40

[218] Even A, Shankaranarayanan G (2009) Dual assessment of data quality in customer databases. Journal of Data and Information Quality (JDIQ) 1(3):15

[219] Even A, Shankaranarayanan G (2009) Utility cost perspectives in data quality management. Journal of Computer Information Systems 50(2)
[220] Even A, Shankaranarayanan G, Berger PD (2007) Economics-driven data management: An application to the design of tabular data sets. Knowledge and Data Engineering, IEEE Transactions on 19(6):818–831
[221] Even A, Kolodner Y, Varshavsky R (2010) Designing business-intelligence tools with value-driven recommendations. In: Global Perspectives on Design Science Research, Springer, pp 286–301
[222] Even A, Shankaranarayanan G, Berger PD (2010) Evaluating a model for cost-effective data quality management in a real-world crm setting. Decision Support Systems 50(1):152–163
[223] Even A, Shankaranarayanan G, Berger PD (2010) Inequality in the utility of customer data: Implications for data management and usage. Journal of Database Marketing & Customer Strategy Management 17(1):19–35
[224] Even A, Shankaranarayanan G, Berger PD (2010) Managing the quality of marketing data: Cost/benefit tradeoffs and optimal configuration. Journal of Interactive Marketing 24:209–221
[225] Eysenbach G (2008) Medicine 2.0: social networking, collaboration, participation, apomediation, and openness. Journal of medical Internet research 10(3)
[226] Falorsi PD, Scannapieco M (eds) (2006) Principi Guida per la Qualità dei Dati Toponomastici nella Pubblica Amministrazione (in Italian). ISTAT, serie Contributi, vol. 12. Available at: http://www.istat.it/dati/pubbsci/contributi/Contr_anno2005.htm
[227] Falorsi PD, Pallara S, Pavone A, Alessandroni A, Massella E, Scannapieco M (2003) Improving the Quality of Toponymic Data in the Italian Public Administration. In: Proc. DQCIS 2003 (ICDT Workshop)
[228] Fan W, Lu H, Madnick S, Cheungd D (2001) Discovering and Reconciling Value Conflicts for Numerical Data Integration. Information Systems 26(8)
[229] Farr JN, Jenkins JJ, Paterson DG (1951) Simplification of flesch reading ease formula. Journal of applied psychology 35(5):333
[230] Fawcett T (2004) Roc graphs: Notes and practical considerations for researchers. Machine learning 31:1–38
[231] Fellbaum C (1999) WordNet. Wiley Online Library
[232] Fellegi IP, Holt D (1976) A systematic approach to automatic edit and imputation. Journal of the American Statistical Association 71(353):17–35
[233] Fellegi IP, Sunter AB (1969) A Theory for Record Linkage. Journal of the American Statistical Association 64
[234] Fiedler K, Kareev Y (2006) Does decision quality (always) increase with the size of information samples? some vicissitudes in applying the law of large numbers. Journal of Experimental Psychology: Learning, Memory, and Cognition 32(4):883

[235] Filin S, Doytsher Y (2000) Detection of corresponding objects in linear-based map conflation. Surveying and land information systems 60(2):117–128

[236] Filin S, Doytsher Y (2000) A linear conflation approach for the integration of photogrammetric information and gis data. International archives of photogrammetry and remote sensing 33(B3/1; PART 3):282–288

[237] Fisher C, Lauria E, Chengalur-Smith S, Wang R (2011) Introduction to information quality. AuthorHouse, Bloomington, IN

[238] Fisher CW, Kingma BR (2001) Criticality of Data Quality as Exemplified in Two Disasters. Information Management 39

[239] Fisher CW, Chengalur-Smith I, Ballou DP (2003) The impact of experience and time on the use of data quality information in decision making. Information Systems Research 14(2):170–188

[240] Fitzpatrick G (2000) Understanding the paper health record in practice: Implications for EHRs. In: HIC'2000 Proceedings of Health Informatics Conference, Adelaide, AU, 2000

[241] Fitzpatrick G, Ellingsen G (2012) A review of 25 years of CSCW research in healthcare: Contributions, challenges and future agendas. Computer Supported Cooperative Work (CSCW) DOI 10.1007/s10606-012-9168-0, URL http://www.springerlink.com/index/10.1007/s10606-012-9168-0

[242] Flemming A (2011) Qualitätsmerkmale von Linked Data-veröffentlichenden Datenquellen. Diplomarbeit (Quality Criteria for Linked Data Sources) https://cs.uwaterloo.ca/~ohartig/files/DiplomarbeitAnnikaFlemming.pdf

[243] Flemming, Annika (accessed 2014) Basel Committee on Banking Supervision http://www.ots.treas.gov

[244] Flesch R (1948) A new readability yardstick. Journal of applied psychology 32(3):221

[245] Fortier MFA, Ziou D, Armenakis C, Wang S (2000) Automated updating of road information from aerial images. In: American Society Photogrammetry and Remote Sensing Conference, pp 16–23

[246] Fowler M (2004) UML Distilled: A Brief Guide to the Standard Object Modeling Language. Pearson Education

[247] Fox S, Jones S (2009) The social life of health information. Washington, DC: Pew Internet & American Life Project pp 2009–12

[248] Francalanci C, Pernici B (2004) Data quality assessment from the user's perspective. In: Proceedings of the 2004 international workshop on Information quality in information systems, ACM, pp 68–73

[249] Frey F, Reilly J, of Technology Image Permanence Institute RI (1999) Digital Imaging for Photographic Collections: Foundations for Technical Standards. Image Permanence Institute, URL http://books.google.it/books?id=75QrAQAAMAAJ

[250] Friedman C, Sideli R (1992) Tolerating spelling errors during patient validation. Computers and Biomedical Research 25(5):486–509

[251] Fung B, Wang K, Chen R, Yu PS (2010) Privacy-preserving data publishing: A survey of recent developments. ACM Computing Surveys (CSUR) 42(4):14

[252] Fürber C, Hepp M (2011) Swiqa - a semantic web information quality assessment framework. In: ECIS

[253] Fuxman A, Fazli E, Miller RJ (2005) ConQuer: Efficient Management of Inconsistent Databases. In: Proc. SIGMOD 2005

[254] Gabay Y, Doytsher Y (2000) Features-an approach to matching lines in partly similar engineering maps. Geomatica 54(3):297–310

[255] Galland A, Abiteboul S, Marian A, Senellart P (2010) Corroborating information from disagreeing views. In: WSDM, pp 131–140

[256] Gallegos I, Gates A, Tweedie C (2010) Dapros: A data property specification tool to capture scientific sensor data properties. In: Proceedings of ER Workshops, Vancouver, BC, Canada, 2010

[257] Gamble M, Goble C (2011) Quality, trust, and utility of scientific data on the web: Towards a joint model. In: ACM WebSci, pp 1–8

[258] Gangadharan GR, Weiss M, D'Andrea V, Iannella R (2007) Service license composition and compatibility analysis. In: ICSOC, pp 257–269

[259] Ge M (2009) Information quality assessment and effects on inventory decision-making. PhD thesis, Dublin City University

[260] Ge M, Helfert M (2006) A framework to assess decision quality using information quality dimensions. In: ICIQ, pp 455–466

[261] Ge M, Helfert M (2013) Impact of information quality on supply chain decisions. Journal of Computer Information Systems 53(4)

[262] Geissbuhler A, Safran C, Buchan I, Bellazzi R, Labkoff S, Eilenberg K, Leese A, Richardson C, Mantas J, Murray P, De Moor G (2013) Trustworthy reuse of health data: A transnational perspective. International Journal of Medical Informatics 82(1):1–9, DOI 10.1016/j.ijmedinf.2012.11.003, URL `http://linkinghub.elsevier.com/retrieve/pii/S138650561200202X`

[263] Getoor L, Machanavajjhala A (2012) Entity resolution: theory, practice & open challenges. Proceedings of the VLDB Endowment 5(12):2018–2019

[264] Gil Y, Artz D (2007) Towards content trust of web resources. Web Semantics 5(4):227 – 239

[265] Gil Y, Ratnakar V (2002) Trusting information sources one citizen at a time. In: ISWC, Springer-Verlag, pp 162 – 176

[266] Gillies A (2000) Assessing and improving the quality of information for health evaluation and promotion. Methods of information in medicine 39(3):208–212

[267] Gissler M, Hemminki J, Teperi J, Merilainen J (1995) Data quality after restructuring a national medical registry. Scand J Soc Med 23:75–80

[268] Glasson M, Trepanier J, Patruno V, Daas P, Skaliotis M, Khan A (2013) What does Big data mean for official statistics? Tech. rep., UNECE, URL http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=77170622
[269] Goiser K, Christen P (2006) Towards automated record linkage. In: Proceedings of the fifth Australasian conference on Data mining and analystics-Volume 61, Australian Computer Society, Inc., pp 23–31
[270] Golbeck J (2004) Inferring reputation on the semantic web. In: WWW
[271] Gonzales RC, Woods R (2008) Digital image processing. Prentice Hall
[272] Gonzalez C, Kasper GM (1997) Animation in user interfaces designed for decision support systems: The effects of image abstraction, transition, and interactivity on decision quality. Decision Sciences 28(4):793–823
[273] Gostojić S, Milosavljević B, Konjović Z (2013) Ontological model of legal norms for creating and using legislation. Computer Science and Information Systems 10(1):151–171
[274] Graesser AC, McNamara DS (2011) Computational analyses of multilevel discourse comprehension. Topics in Cognitive Science 3(2):371–398
[275] Graesser AC, McNamara DS, Louwerse MM (2003) What do readers need to learn in order to process coherence relations in narrative and expository text. Rethinking reading comprehension pp 82–98
[276] Graesser AC, McNamara DS, Louwerse MM, Cai Z (2004) Coh-metrix: Analysis of text on cohesion and language. Behavior Research Methods, Instruments, & Computers 36(2):193–202
[277] Greco G, Lembo D (2004) Data Integration with Preferences Among Sources. In: Proc. ER 2004
[278] Greco G, Greco S, Zumpano E (2003) A Logical Framework for Querying and Repairing Inconsistent Databases. Transactions on Knowledge and Data Engineering 15(6):1389–1408
[279] Gruenheid A, Dong XL, Srivastava D (2014) Incremental record linkage. PVLDB 7(9):697–708
[280] Grünwald PD (2007) The minimum description length principle. MIT press
[281] Gu L, Baxter RA (2004) Adaptive filtering for efficient record linkage. In: SDM, SIAM, pp 477–481
[282] Gu L, Baxter R, Vickers D, Rainsford C (2003) Record Linkage: Current Practice and Future Directions. Technical Report 03/83, CMIS 03/83
[283] Guéret C, Groth P, Stadler C, Lehmann J (2012) Assessing linked data mappings using network measures. In: ESWC
[284] Gunning R (1952) The Technique of Clear Writing. McGraw Hill International Book
[285] Guo S, Dong XL, Srivastava D, Zajac R (2010) Record linkage with uniqueness constraints and erroneous values. Proceedings of the VLDB Endowment 3(1-2):417–428

[286] Guptil C, Morrison J (1995) Elements of Spatial Data Quality. Elsevier Science Ltd, Oxford, UK

[287] H Tang NJ, Kapoor A (2011) Learning a blind measure of perceptual image quality. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 305–312

[288] Hagan MT, Demuth HB, Beale MH, et al (1996) Neural network design, vol 1. Pws Boston

[289] Haines M (2010) Information quality research from the healthcare perspective. In: Proceedings of the Fourth MIT Information Quality Industry Symposium, July 14-16, 2010

[290] Hall PA, Dowling G (1980) Approximate String Comparison. ACM Computing Surveys 12(4):381–402

[291] Hall R, Fienberg SE (2011) Privacy-preserving record linkage. In: Privacy in statistical databases, Springer, pp 269–283

[292] Halliday M, Hasan R (1976) Cohesion in English. English language series, Longman, URL http://books.google.it/books?id=zMBZAAAAMAAJ

[293] Halpin H, Hayes P, McCusker JP, McGuinness D, Thompson HS (2010) When owl:sameas isn't the same: An analysis of identity in linked data. In: Proceedings of the 9th International Semantic Web Conference (ISWC), vol 1, pp 53–59

[294] Hammer M, Champy J (2009) Reengineering the Corporation: Manifesto for Business Revolution, A. Collins Business Essentials, HarperCollins, URL http://books.google.it/books?id=mjvGTXgFl6cC

[295] Han J, Kamber M (2000) Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers

[296] Härle P, Heuser M, Pfetsch S, Poppensieker T (2010) Basel iii. what the draft proposals might mean for european banking. Online verfügbar unter http://www mckinsey com/clientservice/Financial_ Servicvices/-Knowledge_Highlights/˜/media/Reports/Financial_Services/MoCIB 10_Basel3 ashx, zuletzt geprüft am 30:2011

[297] Harper RHR, O'Hara KPA, Sellen AJ, Duthie DJR (1997) Toward the paperless hospital? a case study of document use by anaesthetists. British Journal of Anaesthesia 78:762–767

[298] Harrison MI, Koppel R, Bar-Lev S (2007) Unintended consequences of information technologies in health care - an interactive sociotechnical analysis. Journal of the American Medical Informatics Association 14(5):542–549

[299] Hartig O (2008) Trustworthiness of data on the web. In: STI Berlin and CSW PhD Workshop, Berlin, Germany

[300] Hartig O (2009) Provenance Information in the Web of Data. In: Proceedings of the Linked Data on the Web (LDOW'09), Workshop of the World Wide Web Conference (WWW)

[301] Hartig O, Harth A, Sequeda J (eds) (2010) Proceedings of the First International Workshop on Consuming Linked Data, Shanghai, China,

November 8, 2010, CEUR Workshop Proceedings, vol 665, CEUR-WS.org, URL `http://ceur-ws.org/Vol-665`

[302] Hasler D, Süsstrunk SE (2003) Measuring colorfulness in natural images. Human Vision and Electronic Imaging VIII 5007:87–95

[303] Hassanzadeh O, Chiang F, Lee HC, Miller RJ (2009) Framework for evaluating clustering algorithms in duplicate detection. Proceedings of the VLDB Endowment 2(1):1282–1293

[304] Hastings J (2008) Automated conflation of digital gazetteer data. International Journal of Geographical Information Science 22(10):1109–1127

[305] Hayes P (2004) RDF Semantics. Recommendation, World Wide Web Consortium, `http://www.w3.org/TR/2004/REC-rdf-mt-20040210`

[306] Heath T, Bizer C (2011) Linked Data: Evolving the Web into a Global Data Space. Morgan & Claypool

[307] Heinrich B, Kaiser M, Klier M (2007) How to measure data quality - a metric based approach. In: In appraisal for: International Conference on Information Systems

[308] Heinrich B, Kaiser M, Klier M (2008) Does the eu insurance mediation directive help to improve data quality? - a metric-based analysis. In: Golden W, Acton T, Conbo K, van der Heijden H, Tuunainen V (eds) Conference proceedings / ECIS 2008, 16th European Conference on Information Systems : June 9th - 11th 2008, Galway, Ireland

[309] Heinrich B, Klier M, Kaiser M (2009) A procedure to develop metrics for currency and its application in crm. Journal of Data and Information Quality (JDIQ) 1(1):5

[310] Hernández MA, Stolfo SJ (1995) The merge/purge problem for large databases. In: ACM SIGMOD Record, ACM, vol 24, pp 127–138

[311] Hernandez MA, Stolfo SJ (1998) Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem. Journal of Data Mining and Knowledge Discovery 1(2)

[312] Herzfeld T, Weiss C (2003) Corruption and legal (in) effectiveness: an empirical investigation. European Journal of Political Economy 19(3):621–632

[313] Hinchcliff R, Greenfield D, Moldovan M, Westbrook JI, Pawsey M, Mumford V, Braithwaite J (2012) Narrative synthesis of health service accreditation literature. BMJ Quality & Safety 21(12):979–991, DOI 10.1136/bmjqs-2012-000852, URL `http://qualitysafety.bmj.com/lookup/doi/10.1136/bmjqs-2012-000852`

[314] Hogan A, Harth A, Passant A, Decker S, Polleres A (2010) Weaving the pedantic web. In: LDOW

[315] Hogan A, Umbrich J, Harth A, Cyganiak R, Polleres A, Decker S (2012) An empirical survey of linked data conformance. Journal of Web Semantics

[316] Hogan WR, Wagner MM (1997) Accuracy of data in computer-based patient records. Journal of the American Medical Informatics Association 4(5):342–355

[317] Hopkins D, King G (2010) A method of automated nonparametric content analysis for social science. American Journal of Political Science 54(1):229–247

[318] Hovenga EJ (2010) Health informatics: an overview, vol 151. IOS Press

[319] Hristovski D, Rogac M, Markota M (2000) Using data warehousing and OLAP in public health care. In: Proceedings of the AMIA Symposium, American Medical Informatics Association, p 369

[320] Huaman MA, Araujo-Castillo RV, Soto G, Neyra JM, Quispe JA, Fernandez MF, Mundaca CC, Blazes DL (2009) Impact of two interventions on timeliness and data quality of an electronic disease surveillance system in a resource limited setting (peru): a prospective evaluation. BMC medical informatics and decision making 9(1):16

[321] Huang J, Ertekin S, Giles CL (2006) Efficient name disambiguation for large-scale databases. In: Knowledge Discovery in Databases: PKDD 2006, Springer, pp 536–544

[322] Hwang MI, Lin JW (1999) Information dimension, information overload and decision quality. Journal of Information Science 25(3):213–218

[323] I3A (2007) Fundamentals and review of considered test methods. CPIQ Initiative Phase 1 White Paper

[324] Imatest (2010) Digital Image Quality Testing. http://www.imatest.com

[325] Inan A, Kantarcioglu M, Bertino E, Scannapieco M (2008) A hybrid approach to private record linkage. In: Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on, IEEE, pp 496–505

[326] Inan A, Kantarcioglu M, Ghinita G, Bertino E (2010) Private record matching using differential privacy. In: Proceedings of the 13th International Conference on Extending Database Technology, ACM, pp 123–134

[327] International Conference on Information Quality (IQ/ICIQ) (accessed 2015) `http://www.iqconference.org/`

[328] International Monetary Fund (accessed 2014) `http://dsbb.imf.org/`

[329] International Organization for Standardization (accessed 2014) `http://www.iso.org`

[330] INTERPARES Project (accessed 2014) `http://www.interpares.org`

[331] Iselin ER (1988) The effects of information load and information diversity on decision quality in a structured decision task. Accounting, organizations and Society 13(2):147–164

[332] ISO (2000) Quality management and quality assurance. Vocabulary. ISO 84021994. International Organization for Standardization, 1994

[333] ISO (2005) Image technology colour management - Architecture, profile format and data structure - Part 1: Based on ICC.1:2004-10. ISO 15076-1. ISO, 2005

[334] ISO (accessed February 09, 2012) Information technology – Multimedia content description interface – Part 1: Systems. URL `http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=34228`

[335] ITU (2002) Methodology for the subjective assessment of the quality for television pictures. Tech. rep., ITU-R Rec. BT. 500-11

[336] Jaccard P (1901) Etude comparative de la distribution florale dans une portion des Alpes et du Jura. Impr. Corbaz

[337] Jacobi I, Kagal L, Khandelwal A (2011) Rule-based trust assessment on the semantic web. In: International conference on Rule-based reasoning, programming, and applications series, pp 227 – 241

[338] Jain AK (2001) Corruption: a review. Journal of economic surveys 15(1):71–121

[339] James G, Witten D, Hastie T, Tibshirani R (2013) An Introduction to Statistical Learning with Applications in R. Springer Texts in Statistics

[340] Janssen T (2001) Computational Image Quality. SPIE Press

[341] Janssen T, Blommaert F (2000) A computational approach to image quality. Displays 21:129–142

[342] Jarke M, Lenzerini M, Vassiliou Y, Vassiliadis P (eds) (1995) Fundamentals of Data Warehouses. Springer Verlag

[343] Jarke M, Jeusfeld MA, Quix C, Vassiliadis P (1999) Architecture and Quality in Data Warehouses: an Extended Repository Approach. Information Systems

[344] Jaro MA (1985) Advances in Record Linkage Methodologies as Applied to Matching the 1985 Cencus of Tampa, Florida. Journal of American Statistical Society 84(406):414–420

[345] Jarvenpaa SL, Dickson GW, DeSanctis G (1985) Methodological issues in experimental is research: Experiences and recommendations. MIS quarterly 9(2)

[346] Jeffery S, Alonso M Gand Franklin, Hong W, Widom J (2005) A Pipelined Framework for Online Cleaning of Sensor Data Streams. Tech. rep., Computer Science Division (EECS) University of California, uCB/CSD-5-1413

[347] Jeffery S, Alonso M Gand Franklin, Hong W, Widom J (2005) A Pipelined Framework for Online Cleaning of Sensor Data Streams. Tech. rep., Computer Science Division (EECS) University of California, uCB/CSD-5-1413

[348] Jeffery S, Garofalakis M, Franklin M (2006) Adaptive cleansing for rfid data streams. In: Proceedings of Very Large Database Conference (VLDB 2006), Seoul, Korea, 2006

[349] Jha AK, DesRoches CM, Kralovec PD, Joshi MS (2010) A progress report on electronic health records in U.S. hospitals. Health Affairs 29(10):1951–1957, DOI 10.1377/hlthaff.2010.0502, URL `http://content.healthaffairs.org/cgi/doi/10.1377/hlthaff.2010.0502`

[350] Johnson S, Kaufmann D, Zoido-Lobaton P (1998) Regulatory discretion and the unofficial economy. American Economic Review 88(2):387–392

[351] Jung W (2004) A review of research: an investigation of the impact of data quality on decision performance. In: Proceedings of the 2004 inter-

national symposium on Information and communication technologies, Trinity College Dublin, pp 166–171

[352] Jung W, Olfman L, Ryan T, Park YT (2005) An experimental study of the effects of contextual data quality and task complexity on decision performance. In: Information Reuse and Integration, IEEE Conference, 2005. IRI-2005, IEEE, pp 149–154

[353] Juran J (1988) Juran on planning for quality. The Free Press, Ney York

[354] Kalashnikov DV, Mehrotra S (2006) Domain-independent data cleaning via analysis of entity-relationship graph. ACM Transactions on Database Systems (TODS) 31(2):716–767

[355] Karakasidis A, Verykios VS (2010) Advances in privacy preserving record linkage. E-Activity and Intelligent Web Construction: Effects of Social Design pp 22–29

[356] Kargupta H, Datta S, Wang Q, Sivakumar K (2003) On the privacy preserving properties of random data perturbation techniques. In: Data Mining, 2003. ICDM 2003. Third IEEE International Conference on, IEEE, pp 99–106

[357] Karsh BT, Weinger MB, Abbott PA, Wears RL (2010) Health information technology: fallacies and sober realities. Journal of the American Medical Informatics Association 17(6):617–623

[358] Kay M, Santos J, Takane M (2011) mHealth: new horizons for health through mobile technologies. World Health Organization

[359] Kazley AS, Ozcan YA (2008) Do hospitals with electronic medical records (EMRs) provide higher quality care? an examination of three clinical conditions. Medical Care Research and Review 65(4):496–513

[360] Keelan BW (2002) Handbook of Image Quality: Characterization and Prediction. CRC Press

[361] Keller KL, Staelin R (1987) Effects of quality and quantity of information on decision effectiveness. Journal of consumer research pp 200–213

[362] Kerr K, Norris T (2008) Improving health care data quality: A practitioner's perspective. International Journal of Information Quality 2(1):39, DOI 10.1504/IJIQ.2008.019562, URL `http://www.inderscience.com/link.php?id=19562`

[363] Kerr K, Norris T, Stockdale R (2007) Data quality information and decision making: a healthcare case study. In: 18th Australasian Conference on Information Systems, Toowoomba, pp 5–7

[364] Kerr KA, Norris T, Stockdale R (2008) The strategic management of data quality in healthcare. Health Informatics Journal 14(4):259–266, DOI 10.1177/1460458208096555, URL `http://jhi.sagepub.com/cgi/doi/10.1177/1460458208096555`

[365] Keßler C, Janowicz K, Bishr M (2009) An agenda for the next generation gazetteer: Geographic information contribution and retrieval. In: Proceedings of the 17th ACM SIGSPATIAL international conference on advances in Geographic Information Systems, ACM, pp 91–100

[366] Kettinger W, Grover V (1995) Special section: Toward a theory of business process change management. Journal of Management Information Systems 12(1):9–30

[367] Kim W, Seo J (1991) Classifying Schematic and Data Heterogeneity in Multidatabase Systems. IEEE Computer 24(12):12–18

[368] Kimball R (1998) The data warehouse lifecycle toolkit: expert methods for designing, developing, and deploying data warehouses. John Wiley & Sons

[369] Kincaid JP, Fishburne Jr RP, Rogers RL, Chissom BS (1975) Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Tech. rep., DTIC Document

[370] Kitson HD (1921) The mind of the buyer: A psychology of selling, vol 21549. Macmillan

[371] Klare GR (1974) Assessing readability. Reading research quarterly pp 62–102

[372] Klein A, Lehner W (2009) Representing data quality in sensor data streaming environments. Journal of Data and Information Quality 1(2)

[373] Koda K (2005) Insights into second language reading: A cross-linguistic approach. Cambridge University Press

[374] Kohn LT, Corrigan JM, Donaldson MS (eds) (2000) To Err Is Human: Building a Safer Health System. Institute of Medicine (IOM)

[375] Kolodner Y (2009) Enhancing business-intelligence tools with value-driven recommendations. PhD thesis, BEN-GURION UNIVERSITY OF THE NEGEV

[376] Kolodner Y, Even A (2009) Integrating value-driven feedback and recommendation mechanisms into business intelligence systems. In: ECIS, pp 1987–1998

[377] Königer P, Janowitz K (1995) Drowning in information, but thirsty for knowledge. International Journal of Information Management 15(1):5–16

[378] Köpcke H, Rahm E (2010) Frameworks for entity matching: A comparison. Data & Knowledge Engineering 69(2):197–210

[379] Köpcke H, Thor A, Rahm E (2010) Evaluation of entity resolution approaches on real-world match problems. Proceedings of the VLDB Endowment 3(1-2):484–493

[380] Koppel R, Metlay JP, Cohen A, Abaluck B, Localio AR, Kimmel SE, Strom BL (2005) Role of computerized physician order entry systems in facilitating medication errors. Journal of the American Medical Association 293:1197–âĂŞ1203

[381] Krawczyk H, Wiszniewski B (2003) Visual GQM Approach to Quality-driven Development of Electronic Documents. In: Proc. 2nd International Workshop on Web Document Analysis (WDA2003)

[382] Krötzsch M, Speiser S (2011) Sharealike your data: Self-referential usage policies for the semantic web. In: International Semantic Web Conference (1), pp 354–369

[383] Kukich K (1992) Techniques for automatically correcting words in text. ACM Computing Surveys (CSUR) 24(4):377–439

[384] Kusuma T, Zepernick HJ (2003) A reduced-reference perceptual quality metric for in-service image quality assessment. In: Mobile Future and Symposium on Trends in Communications, 2003. SympoTIC '03. Joint First Workshop on, pp 71 – 74

[385] Lafferty JD, McCallum A, Pereira FCN (2001) Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, ICML '01, pp 282–289, URL http://dl.acm.org/citation.cfm?id=645530.655813

[386] Lait A, Randell B (1996) An assessment of name matching algorithms. Technical Report Series-University of Newcastle Upon Tyne Computing Science

[387] Lantz B (2013) Machine Learning with R. Packt Publishing Ltd

[388] Lapointe L (2006) Getting physicians to accept new information technology: insights from case studies. Canadian Medical Association Journal 174(11):1573–1578, DOI 10.1503/cmaj.050281, URL http://www.cmaj.ca/cgi/doi/10.1503/cmaj.050281

[389] Larsen MD, Rubin DB (1989) An Iterative Automated Record Matching using Mixture Models. Journal of American Statistical Association 79:32–41

[390] Larson EC, Chandler DM (2010) Most apparent distortion: full-reference image quality assessment and the role of strategy. J Electronic Imaging 19(1):011,006–1–011,006–21)

[391] Lee BK, Lee WN (2004) The effect of information overload on consumer choice quality in an on-line environment. Psychology & Marketing 21(3):159–183

[392] Lee C, Rey T, Mentele J, Garver M (2005) Structured neural network techniques for modeling loyalty and profitability. Proceedings of SAS User Group International (SUGI 30) pp 082–30

[393] Lee YW, Strong DM, Kahn BK, Wang RY (2001) AIMQ: A Methodology for Information Quality Assessment. Information and Management

[394] Lehmann J, Gerber D, Morsey M, Ngonga Ngomo AC (2012) DeFacto - Deep Fact Validation. In: ISWC, Springer Berlin / Heidelberg

[395] Lehti P, Fankhauser P (2005) Probabilistic iterative duplicate detection. In: OTM Conferences (2), pp 1225–1242

[396] Lei Y, Uren V, Motta E (2007) A framework for evaluating semantic metadata. In: 4th International Conference on Knowledge Capture, ACM, no. 8 in K-CAP '07, pp 135 – 142

[397] Leiner F, Gaus W, Haux R, Leiner F, Gaus W, Haux R (2003) Medical data management. Springer

[398] Lenzerini M (2002) Data Integration: A Theoretical Perspective. In: Proc. PODS 2002

[399] Letzring TD, Wells SM, Funder DC (2006) Information quantity and quality affect the realistic accuracy of personality judgment. Journal of personality and social psychology 91(1):111

[400] Levy AY, Mendelzon AO, Sagiv Y, Srivastava D (1995) Answering Queries Using Views. In: Proc. PODS 1995

[401] Li L, Goodchild M (2012) Automatically and accurately matching objects in geospatial datasets. Adv Geo-Spat Inf Sci 10:71–79

[402] Li X, Dong XL, Lyons K, Srivastava D (1999) Truth Finding on the Deep Web: Is the Problem Solved? PVLDB

[403] Liaw S, Chen H, Maneze D, Taggart J, Dennis S, Vagholkar S, Bunker J (2011) Health reform: is current electronic information fit for purpose. Emergency Medicine Australasia

[404] Liaw S, Rahimi A, Ray P, Taggart J, Dennis S, de Lusignan S, Jalaludin B, Yeo A, Talaei-Khoei A (2013) Towards an ontology for data quality in integrated chronic disease management: A realist review of the literature. International Journal of Medical Informatics 82(1):10–24, DOI 10.1016/j.ijmedinf.2012.10.001, URL `http://linkinghub.elsevier.com/retrieve/pii/S1386505612001931`

[405] Liaw ST, Taggart J, Dennis S, Yeo A (2011) Data quality and fitness for purpose of routinely collected dataâĂŞa general practice case study from an electronic practice-based research network (ePBRN). In: AMIA Annual Symposium Proceedings, American Medical Informatics Association, vol 2011, p 785

[406] Lim EP, Chiang RH (Singapore, Singapore, 1998) A Global Object Model for Accommodating Instance Heterogeneities. In: Proc. ER'98

[407] Lin J, Mendelzon AO (1998) Merging Databases Under Constraints. International Journal of Cooperative Information Systems 7(1):55–76

[408] Linked Open Data (LOD) (2006) `http://linkeddata.org/`

[409] Liu L, Chi L (2002) Evolutionary data quality. In: 7th International Conference on Information Quality, Boston, MA, USA

[410] LIVE video (2009) Live video quality database. URL `http://live.ece.utexas.edu/research/quality/live_video.html`

[411] Lohningen H (1999) Teach Me Data Analysis. Springer-Verlag

[412] Lorence D, Jameson R (2001) Adoption of information quality practices in US healthcare organisations. a national assessment. International Journal of Quality and Reliability Management 19(6):737–756

[413] Lorence DP (2003) The perils of data misreporting. Communications of the ACM 46(11):85–88

[414] Loshin D (2004) Enterprise Knowledge Management - The Data Quality Approach. Morgan Kaufmann Series in Data Management Systems

[415] Loshin D (2011) The practitioner's guide to data quality improvement. Morgan Kaufmann, Burlington, MA

[416] Lovern E (2000) Accreditation gains attention. Modern healthcare 30(47):46

[417] Low W, Lee M, Ling T (2001) A Knowledge-based Approach for Duplicate Elimination in Data Cleaning. Information Systems 26(8)

[418] Lundstrom C (2006) Technical report: Measuring digital image quality. Tech. rep., Linkoping UniversityLinkoping University, Visual Information Technology and Applications (VITA), The Institute of Technology

[419] Lupo C, Batini C (2003) A federative approach to laws access by citizens: The "normeinrete" system. In: Electronic Government, Springer, pp 413–416

[420] Lupo C, De Santis L, Batini C (2005) Legalurn: a framework for organizing and surfing legal documents on the web. In: Challenges of Expanding Internet: E-Commerce, E-Business, and E-Government, Springer, pp 313–327

[421] de Lusignan S, Stephens PN, Adal N, Majeed A (2002) Does feedback improve the quality of computerized medical records in primary care? Journal of the American Medical Informatics Association 9(4):395–401

[422] de Lusignan S, Khunti K, Belsey J, Hattersley A, van Vlymen J, Gallagher H, Millett C, Hague NJ, Tomson C, Harris K, Majeed A (2010) A method of identifying and correcting miscoding, misclassification and misdiagnosis in diabetes: a pilot and validation study of routinely collected data. Diabetic Medicine 27(2):203–209, DOI 10.1111/j.1464-5491.2009.02917.x, URL `http://doi.wiley.com/10.1111/j.1464-5491.2009.02917.x`

[423] MacDonald L, Jacobson R (2006) Assessing image quality. In Digital heritage: applying digital imaging to cultural heritage. Elsevier Butterworth-Heinemann

[424] Madhavan J, Ko D, Kot L, Ganapathy V, Rasmussen A, Halevy AY (2008) Google's Deep Web crawl. PVLDB 1(2):1241–1252

[425] Malin B (2005) Unsupervised name disambiguation via social network similarity. In: Workshop on link analysis, counterterrorism, and security, vol 1401, pp 93–102

[426] Mann R, Williams J (2003) Standards in medical record keeping. Clinical Medicine, Journal of the Royal College of Physicians 3(4):329–332

[427] Mann WC, Thompson SA (1988) Rhetorical structure theory: Toward a functional theory of text organization. Text 8(3):243–281

[428] Manzoor A, Truong H, S D (2008) On the evaluation of quality of context. In: European Conference on Smart Sensing & Context (EuroSSC), Zurich, Switzerland, 2008

[429] Martinez A, Hammer J (2005) Making quality count in biological data sources. In: IQIS '05: Proceedings of the 2nd international workshop on Information quality in information systems, ACM Press, New York, NY, USA

[430] Marziliano P, Dufaux F, Winkler S, Ebrahimi T (2002) A no-reference perceptual blur metric. In: IEEE 2002 International Conference on Image Processing, pp 57–60
[431] Maydanchik A (2007) Data quality assessment. Data quality for practitioners series, Technics Publications, Bradley Beach, NJ
[432] McKeon A (2003) Barclays bank case study: Using artificial intelligence to benchmark organizational data flow quality. In: Proceeding of the Eighth International Conference on Information Quality
[433] McKinsey Global Institute (2013) Open data: Unlocking innovation and performance with liquid information
[434] McLaughlin GH (1969) Smog grading: A new readability formula. Journal of reading 12(8):639–646
[435] McNamara DS, Louwerse MM, Graesser AC (2002) Coh-metrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension. Unpublished Grant proposal, University of Memphis, Memphis, Tennessee
[436] Memorandum S (accessed 2014) `http://epp.eurostat.ec.europa.eu/portal/page/portal/pgp\_ess/0\_DOCS/estat/SCHEVENINGEN\_MEMORANDUM\%20Final\%20version\_0.pdf`
[437] Mendes P, Mühleisen H, Bizer C (2012) Sieve: Linked data quality assessment and fusion. In: LWDM
[438] Michelson M, Knoblock CA (2006) Learning blocking schemes for record linkage. Proceedings of the National Conference on Artificial Intelligence 21(1):440
[439] Michelson M, Knoblock CA (2007) Mining heterogeneous transformations for record linkage. In: Proceedings of the 6th International Workshop on Information Integration on the Web, pp 68–73
[440] Mikkelsen G, Aasly J (2005) Consequences of impaired data quality on information retrieval in electronic patient records. International journal of medical informatics 74(5):387–394
[441] Minami M, et al (2002) Using arcmap. In: Using ArcMap, ESRI
[442] Minkov E, Cohen WW, Ng AY (2006) Contextual search and name disambiguation in email using graphs. In: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, pp 27–34
[443] Minton SN, Nanjo C, Knoblock CA, Michalowski M, Michelson M (2005) A heterogeneous field matching method for record linkage. In: Data Mining, Fifth IEEE International Conference on, IEEE, pp 8–pp
[444] Missier P, Batini C (2003) A Model for Information Quality Management Framework for Cooperative Information Systems. In: Proc. 11th Italian Symposium on Advanced Database Systems (SEDB 2003)
[445] Missier P, Batini C (2003) An Information Quality Management Framework for Cooperative Information Systems. In: Proc. International Conference on Information Systems and Engineering (ISE 2003)

[446] Missier P, Batini C (2003) A multidimensional model for information quality in cooperative information systems. In: Proceedings of the 8th International conference on Information quality

[447] Missier P, Lack G, Verykios V, Grillo F, Lorusso T, Angeletti P (2003) Improving Data Quality in Practice: a Case Study in the Italian Public Administration. Parallel and Distributed Databases 13(2):135–160

[448] Mitchell J, Westerduin F (2008) Emergency department information system diagnosis: how accurate is it? Emergency Medicine Journal 25(11):784–784

[449] Monge A, Elkan C (Tucson, AZ, 1997) An Efficient Domain Independent Algorithm for Detecting Approximate Duplicate Database Records. In: Proc. SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'97)

[450] Moody DL, Walsh P (1999) Measuring the value of information-an asset valuation approach. In: ECIS, pp 496–512

[451] Moorthy A, Bovik A (2011) Visual quality assessment algorithms: what does the future hold? Multimedia Tools and Applications 51:675–696

[452] Morbey G (2013) Data Quality for Decision Makers: A dialog between a board member and a DQ expert. Springer

[453] Mostafavi M, G E, Jeansoulin R (2004) Ontology-based method for quality assessment of spatial data bases. In: International Symposium on Spatial Data Quality, vol 4, pp 49–66

[454] Motik B, Patel-Schneider PF, Parsia B, Bock C, Fokoue A, Haase P, Hoekstra R, Horrocks I, Ruttenberg A, Sattler U, Smith M (2008) OWL 2 web ontology language: Structural specification and functional-style syntax. Last call working draft, W3C, `http://www.w3.org/2007/OWL/draft/owl2-syntax/`

[455] Motro A, Anokhin P (2005) Fusionplex: Resolution of Data Inconsistencies in the Data Integration of Heterogeneous Information Sources. Information Fusion

[456] Motro A, Ragov I (1998) Estimating Quality of Databases. In: Proc. 3rd International Conference on Flexible Query Answering Systems (FQAS'98)

[457] Murero M, Rice RE (2013) The Internet and health care: theory, research, and practice. Routledge

[458] Musavi MT, Shirvaikar MV, Ramanathan E, Nekovei A (1988) A vision based method to automate map processing. Pattern Recognition 21(4):319–326

[459] Muthu S, Withman L, Cheraghi S (1999) Business Process Reengineering: a Consolidated Methodology. In: Proc. 4th Annual International Conference on Industrial Engineering Theory, Applications and Practice

[460] Nahm ML, Pieper CF, Cunningham MM (2008) Quantifying data quality for clinical trials using electronic data capture. PLoS ONE

3(8):e3049, DOI 10.1371/journal.pone.0003049, URL http://dx.plos.org/10.1371/journal.pone.0003049

[461] NASSCOM (2012) Big Data-The Next Big Thing. URL http://www.nasscom.in/sites/default/files/researchreports/softcopy/Big\%20Data\%20Report\%202012.pdf

[462] Naumann F (2002) Quality-Driven Query Answering for Integrated Information Systems, Lecture Notes in Computer Science, vol 2261. Springer-Verlag

[463] Naumann F, Häussler M (2002) Declarative Data Merging with Conflict Resolution. In: 7th International Conference on Information Quality

[464] Naumann F, Leser U, Freytag JC (1999) Quality-driven Integration of Heterogenous Information Systems. In: Proc. VLDB'99

[465] Naumann F, Freytag JC, Leser U (2004) Completeness of Integrated Information Sources. Information Systems 29(7):583–615

[466] Navarro G (2001) A Guided Tour of Approximate String Matching. ACM Computing Surveys 31:31–88

[467] Ndabarora E, Chipps JA, Uys L (2014) Systematic review of health data quality management and best practices at community and district levels in LMIC. Information Development 30(2):103–120, DOI 10.1177/0266666913477430, URL http://idv.sagepub.com/cgi/doi/10.1177/0266666913477430

[468] Nebel B, Lakemeyer G (eds) (1994) Foundations of Knowledge Representation and Reasoning, vol 810, lecture notes in artificial intelligence edn. Springer-Verlag

[469] Neely MP, Cook JS (2011) Fifteen years of data and information quality literature: Developing a research agenda for accounting. Journal of Information Systems 25(1):79–108

[470] Newbold N, Gillam L (2010) The linguistics of readability: the next step for word processing. In: Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics and Writing: Writing Processes and Authoring Aids, Association for Computational Linguistics, pp 65–72

[471] Newcombe HB, Kennedy JM, Axford SJ, James APF (1959) Automatic Linkage of Vital Records. Science 130

[472] Nicholson R, Penney D (2004) Quality data critical to healthcare decision making. Proceedings of the 2004 American Health Information Management Association Chicago, IL: American Health Information Management Association

[473] Nigam K, McCallum A, Thrun S, Mitchell T (2000) Text Classification from Labeled and Unlabeled Documents using EM. Machine Learning 39:103–134

[474] Nin J, Muntes-Mulero V, Martinez-Bazan N, Larriba-Pey JL (2007) On the use of semantic blocking techniques for data cleansing and integration. In: Database Engineering and Applications Symposium, 2007. IDEAS 2007. 11th International, IEEE, pp 190–198

[475] Nishiyama M, Okabe T, Sato I, Sato Y (2011) Aesthetic quality classification of photographs based on color harmony. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pp 33–40

[476] Nov O, Schecter W (2012) Dispositional resistance to change and hospital physicians' use of electronic medical records: A multidimensional perspective: Journal of the american society for information science and technology. Journal of the American Society for Information Science and Technology 63(4):648–656, DOI 10.1002/asi.22602, URL `http://doi.wiley.com/10.1002/asi.22602`

[477] Nuray-Turan R, Kalashnikov DV, Mehrotra S (2013) Adaptive connection strength models for relationship-based entity resolution. Journal of Data and Information Quality (JDIQ) 4(2):8

[478] Object Management Group (OMG) (2003) Unified Modeling Language Specification, Version 1.5

[479] Office of Management and Budget (2002) Information Quality Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Agencies. `http://www.whitehouse.gov/omb/fedreg/reproducible.html`

[480] Olson JE (2003) Data quality: the accuracy dimension. Morgan Kaufmann

[481] Ong MS, Coiera E (2010) Safety through redundancy: a case study of in-hospital patient transfers. Quality and Safety in Health Care 19(5):1–7

[482] ORACLE (accessed 2014) `http://www.oracle.com/solutions/business-intelligence`

[483] Orfanidis L, Bamidis PD, Eaglestone B (2004) Data quality issues in electronic health records: an adaptation framework for the greek health system. Health informatics journal 10(1):23–36

[484] Organization for Economic Co-Operation and Development (1994) Improving the quality of laws and regulations

[485] Osservatorio Legislativo Interregionale, Italy (2007) Regole e suggerimenti per la redazione di testi normativi, in italian

[486] Ostman A (1997) The Specifications and Evaluation of Spatial Data Quality. In: Proc. 18th ICA/ACI International Conference

[487] Ozsu T, Valduriez P (2000) Principles of Distributed Database Systems. Prentice Hall

[488] Ozuru Y, Dempsey K, McNamara DS (2009) Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. Learning and instruction 19(3):228–242

[489] Papadimitriou CH (2003) Computational complexity. John Wiley and Sons Ltd.

[490] Papakonstantinou Y, Abiteboul S, Garcia-Molina H (1996) Object Fusion in Mediator Systems. In: Proc. VLDB 1996

[491] Parssian A, Sarkar S, Jacob V (1999) Assessing Data Quality for Information Products. In: Proc. 20th International Conference on Information Systems (ICIS 99)
[492] Parssian A, Sarkar S, Jacob V (2002) Assessing Information Quality for the Composite Relational Operation Join. In: Proc. 7th International Conference on Information Quality (IQ 2002)
[493] Parssian A, Sarkar S, Jacob V (2004) Assessing Data Quality for Information Products: Impact of Selection, Projection, and Cartesian Product. Management Science 50(7)
[494] Payne RS, McVay S (1971) Songs of humpback whales. Science 173(3997):585–597
[495] Pearl J (1986) Fusion, propagation, and structuring in belief networks. Artificial intelligence 29(3):241–288
[496] Pei L, Dong XL, Maurino M, Srivastava D (2011) Linking temporal records. Frontiers of Computer Science
[497] Perkowitz M, Etzioni O (2000) Adaptive Web-Sites. Communication of the ACM 43(8)
[498] Pernici B, Scannapieco M (2003) Data Quality in Web Information Systems. Journal of Data Semantics
[499] Pessoa A, Falcao A, e Silva A, Nishihara R, Lotufo R (1998) Video quality assessment using objective parameters based on image segmentation. In: Telecommunications Symposium, 1998. ITS '98 Proceedings. SBT/IEEE International, vol 2, pp 498–503
[500] Phua C, Smith-Miles K, Lee V, Gayler R (2012) Resilient identity crime detection. Knowledge and Data Engineering, IEEE Transactions on 24(3):533–546
[501] Pierce E (2002) Extending ip-maps: Incorporating the event driven process chain methodology. In: Proceedings of the International COnference on Information Quality
[502] Pinson M, Wolf S (2004) A new standardized method for objectively measuring video quality. Broadcasting, IEEE Transactions on 50(3):312–322
[503] Pipino L, Lee Y (2011) Medical errors and information quality: A review and research agenda. In: AMCIS'11: Proceedings of the Seventeenth Americas Conference on Information Systems, Detroit, Michigan August 4th-7th 2011
[504] Pipino LL, Lee YW, Wang RY (2002) Data Quality Assessment. Communications of the ACM 45(4)
[505] Pixton B, Giraud-Carrier C (2006) Using structured neural networks for record linkage. In: Proceedings of the Sixth Annual Workshop on Technology for Family History and Genealogical Research
[506] Plebani M (2009) Exploring the iceberg of errors in laboratory medicine. Clinica Chimica Acta 404(1):16–23
[507] Poirier C (Rome, Italy, 2-4 June 1999) A Functional Evaluation of Edit and Imputation Tools. In: UN/ECE Work Statistical Data Editing

[508] Ponomarenko N, Lukin V, Zelensky A, Egiazarian K, Astola J, Carli M, Battisti F (2009) A database for evaluation of full reference visual quality assessment metrics. Advances of Modern Radioelectronics 10:30–45

[509] Porat MU (1977) The Information Economy: Definition and Measurement. ERIC

[510] Porcheret M (2003) Data quality of general practice electronic health records: The impact of a program of assessments, feedback, and training. Journal of the American Medical Informatics Association 11(1):78–86, DOI 10.1197/jamia.M1362, URL http://www.jamia.org/cgi/doi/10.1197/jamia.M1362

[511] Porter EH, Winkler WE, et al (1997) Approximate string comparison and its effect on an advanced record linkage system. In: Advanced record linkage system. US Bureau of the Census, Research Report, Citeseer

[512] Quality of laws Institute (accessed 2014) URL http://www.qualityoflaws.com

[513] Quan H, Li B, Duncan Saunders L, Parsons GA, Nilsson CI, Alibhai A, Ghali WA (2008) Assessing validity of ICD-9-CM and ICD-10 administrative data in recording clinical conditions in a unique dually coded database. Health services research 43(4):1424–1441

[514] Raghunathan S (1999) Impact of information quality and decision-maker quality on decision quality: a theoretical model and simulation analysis. Decision Support Systems 26(4):275–286

[515] Rahimi B, Vimarlund V (2007) Methods to evaluate health information systems in healthcare settings: a literature review. Journal of medical systems 31(5):397–432, PMID: 17918694

[516] Rao J, Doraiswamy S, Thakkar H, Colby L (2006) A deferred cleansing method for rfid data analytics. In: Proceedings of Very Large Database Conference (VLDB 2006), Seoul, Korea, 2006

[517] Recchia G, Louwerse M (2013) A comparison of string similarity measures for toponym matching. In: Proceedings of The First ACM SIGSPATIAL International Workshop on Computational Models of Place, pp 54–61

[518] Redman TC (1996) Data Quality for the Information Age. Artech House

[519] Redman TC (1998) The Impact of Poor Data Quality on the Typical Enterprise. Communications of the ACM

[520] Redman TC (2001) Data Quality The Field Guide. The Digital Press

[521] Renkema J (2001) Undercover research into text quality as a tool for communication management. Reading and writing public documents: problems, solutions and characteristics John Benjamins Publishing Company: Amsterdam pp 37–57

[522] Reuther P, Walter B (2006) Survey on test collections and techniques for personal name matching. International Journal of Metadata, Semantics and Ontologies 1(2):89–99
[523] Reynolds T, Painter I, Streichert L (2013) Data quality: A systematic review of the biosurveillance literature. Online Journal of Public Health Informatics 5(1)
[524] Richards H, White N (2013) Ensuring the quality of health information: The canadian experience. In: Handbook of Data Quality, Springer Berlin Heidelberg, pp 321–346
[525] Richardson M, Domingos P (2006) Markov logic networks. Machine learning 62(1-2):107–136
[526] de Ridder H, Endrikhovski S (2002) Image quality is fun: Reflections on fidelity, usefulness and naturalness. SID Symposium Digest of Technical Papers 33:986–989
[527] Rigby M, Roberts R, Williams J, Clark J, Savill A, Lervy B, Mooney G (1998) Integrated record keeping as an essential aspect of a primary care led health service. BMJ: British Medical Journal 317(7158):579
[528] Risk A, Dzenowagis J (2001) Review of internet health information quality initiatives. Journal of medical Internet research 3(4)
[529] Rittel HWJ, Webber MM (1973) Dilemmas in a general theory of planning. Policy Sciences 4(2):155–169
[530] Rouse DM, Hemami SS (2008) Analyzing the role of visual structure in the recognition of natural image content with multi-scale SSIM. Proc SPIE: HVEI XIII 6806
[531] Ruibin G, Tony K (2006) Syllable alignment: A novel model for phonetic string search. IEICE transactions on information and systems 89(1):332–339
[532] Rula A, Panziera L, Palmonari M, Maurino A (2014) Capturing the currency of dbpedia descriptions and get insight into their validity. In: Proceedings of the 5th International Workshop on Consuming Linked Data (COLD 2014) at the 13th International Semantic Web Conference (ISWC)
[533] Saalfeld AJ (1993) Conflation: Automated map compilation. PhD thesis, University of Maryland at College Park, College Park, MD, USA, uMI Order No. GAX93-27487
[534] Saaty TL (1980) The Analytic Hierarchy Process. McGraw-Hill
[535] Sadinle M, Fienberg SE (2013) A generalized fellegi–sunter framework for multiple record linkage with application to homicide record systems. Journal of the American Statistical Association 108(502):385–397
[536] Sadiq S (ed) (2013) Handbook of Data Quality. Springer Berlin Heidelberg, Berlin, Heidelberg, URL `http://link.springer.com/10.1007/978-3-642-36257-6`
[537] Safra E, Kanza Y, Sagiv Y, Doytsher Y (2006) Efficient integration of road maps. In: Proceedings of the 14th annual ACM international

symposium on Advances in geographic information systems, ACM, pp 59–66

[538] Safra E, Kanza Y, Sagiv Y, Beeri C, Doytsher Y (2010) Location-based algorithms for finding sets of corresponding objects over several geospatial data sets. International Journal of Geographical Information Science 24(1):69–106

[539] Safra E, Kanza Y, Sagiv Y, Doytsher Y (2013) Ad hoc matching of vectorial road networks. International Journal of Geographical Information Science 27(1):114–153

[540] Saha S, Vemuri R (2000) An analysis on the effect of image activity on lossy coding performance. In: Circuits and Systems, 2000. Proceedings. ISCAS 2000 Geneva. The 2000 IEEE International Symposium on, vol 3, pp 295 –298 vol.3

[541] Sala M (2006) Versions of the constitution for europe: Linguistic, textual and pragmatic aspects. Linguistica e filologia 22:139–167

[542] Salamone S, Scannapieco, Scarno M (2014) Web scraping and web mining: new tools for official statistics. In: Proceedings of Societa Italiana di Statistica (SIS 2014), Cagliari, Sardegna, Italy, 2014

[543] Salzberg SL (1997) On comparing classifiers: Pitfalls to avoid and a recommended approach. Data mining and knowledge discovery 1(3):317–328

[544] Sarawagi S, Bhamidipaty A (eds) (Edmonton, Alberta, Canada, 2002) Interactive Deduplication Using Active Learning

[545] Sazzad Z, Kawayoke Y, Horita Y (2000) Mict image quality evaluation database. `http://mict.eng.u-toyama.ac.jp/mict/index2.html`

[546] Scannapieco M, Batini C (2004) Completeness in the Relational Model: A Comprehensive Framework. In: Proc. 9th International Conference on Information Quality (IQ 2004)

[547] Scannapieco M, Virgillito A, Marchetti C, Mecella M, Baldoni R (2004) The DaQuinCIS Architecture: a Platform for Exchanging and Improving Data Quality in Cooperative Information Systems. Information Systems 29(7):551–582

[548] Scannapieco M, Pernici B, Pierce EM (2005) IP-UML: A Methodology for Quality Improvement based on IP-MAP and UML. In: Wang RY, Pierce EM, Madnick SE, Fisher CW (eds) Advances in Management Information Systems - Information Quality (AMIS-IQ) Monograph, Sharpe, M.E.

[549] Scannapieco M, Figotin I, Bertino E, Elmagarmid AK (2007) Privacy preserving schema and data matching. In: Proceedings of the 2007 ACM SIGMOD international conference on Management of data, ACM, pp 653–664

[550] Scannapieco M, Virgillito A, Zardetto D (2013) Placing big data in official statistics: A big challenge? In: Proceedings of 2013 New Techniques and Tools for Statistics (NTTS) Conference, Brussels, Belgium, 2013

[551] Schallehn E, Sattler KU, Saake G (San Jose, CA, 2002) Extensible and Similarity-Based Grouping for Data Integration. In: Proc. of the ICDE 2002

[552] Schapire WWCRE, Singer Y (1998) Learning to order things. In: Advances in Neural Information Processing Systems 10: Proceedings of the 1997 Conference, MIT Press, vol 10, p 451

[553] Schettini R, Gasparini F (2009) A review of redeye detection and removal in digital images through patents. Recent Patents on Electrical Engineering 2(1):45 – 53

[554] Schneier B (2007) Applied cryptography: protocols, algorithms, and source code in C. john wiley & sons

[555] Schober D, Barry S, Lewis ES, Kusnierczyk W, Lomax J, Mungall C, Taylor FC, Rocca-Serra P, Sansone SA (2009) Survey-based naming conventions for use in OBO foundry ontology development. BMC Bioinformatics 10(125)

[556] Sebastian-Coleman L (2013) Measuring data quality for ongoing improvement: a data quality assessment framework. Morgan Kaufmann, Waltham, MA

[557] Sehgal V, Getoor L, Viechnicki PD (2006) Entity resolution in geospatial data integration. In: Proceedings of the 14th annual ACM international symposium on Advances in geographic information systems, ACM, pp 83–90

[558] Senate M (2010) Legislative research and drafting manual

[559] Senter R, Smith E (1967) Automated readability index. Tech. rep., DTIC Document

[560] Seshadrinathan K, Bovik AC (2010) Motion tuned spatio-temporal quality assessment of natural videos. Transaction Imgage Processing 19(2):335–350

[561] Sha K, Shi W (2008) Consistency-driven data quality management of networked sensor systems. Journal of parallel and Distributed Computing 68(9)

[562] Shankaranarayan G, Wang R, Ziad M (2000) Modeling the manufacture of an information product with IP-MAP. In: Proceedings of the 5th International Conference on Information Quality (ICIQ'00), Boston, MA, USA

[563] Shankaranarayanan G, Cai Y (2006) Supporting data quality management in decision-making. Decision Support Systems 42(1):302–317

[564] Sharma G (2002) Digital Color Imaging Handbook. CRC Press, Inc., Boca Raton, FL, USA

[565] Shearer C (2000) The crisp-dm model: the new blueprint for data mining. Journal of data warehousing 5(4):13–22

[566] Sheikh H, Bovik A (2006) Image information and visual quality. Image Processing, IEEE Transactions on 15(2):430 –444

[567] Sheikh HR, Wang Z, Cormack L, Bovik AC (2005) LIVE Image Quality Assessment Database Release 2

[568] Shekarpour S, Katebi S (2010) Modeling and evaluation of trust with an extension in semantic web. Web Semantics: Science, Services and Agents on the World Wide Web 8(1):26 – 36
[569] Shekhar S, Xiong H (2008) Encyclopedia of GIS. Springer
[570] Sheng Y (2003) Exploring the mediating and moderating effects of information quality on firm's endevour on information systems. In: Proceedings of the Eight International Conference on Information Quality 2003 (ICIQ03), Boston, MA, USA, pp 344–352
[571] Sheng Y, Mykytyn P (2002) Information technology investment and firm performance: A perspective of data quality. In: Proceedings of the Seventh International Conference on Information Quality 2002 (ICIQ02), Washington, USA, pp 132–141
[572] Shi W, Fisher P, Goodchild MF (2003) Spatial data quality. CRC Press
[573] Shields MD (1983) Effects of information supply and demand on judgment accuracy: evidence from corporate managers. Accounting Review pp 284–303
[574] Shortell SM, Jones RH, Rademaker AW, Gillies RR, Dranove DS, Hughes EF, Budetti PP, Reynolds KS, Huang CF (2000) Assessing the impact of total quality management and organizational culture on multiple outcomes of care for coronary artery bypass graft surgery patients. Medical care 38(2):207–217
[575] Shortliffe EH, Barnett GO (2001) Medical data: their acquisition, storage, and use. In: Medical Informatics, Springer, pp 41–75
[576] Siau K, Shen Z (2006) Mobile healthcare informatics. Informatics for Health and Social Care 31(2):89–99
[577] Sikora T (2001) The MPEG-7 visual standard for content description-an overview. IEEE Transactions on Circuits and Systems for Video Technology 11(6):696–702
[578] Simnett R (1996) The effect of information selection, information processing and task complexity on predictive accuracy of auditors. Accounting, Organizations and Society 21(7):699–719
[579] Singla P, Domingos P (2006) Entity resolution with markov logic. In: Data Mining, 2006. ICDM'06. Sixth International Conference on, IEEE, pp 572–582
[580] Singleton P, Pagliari C, Detmer D (2009) Critical issues for electronic health records: Considerations from an expert workshop. Tech. rep., Nuffield Trust, UK
[581] Smart PD, Jones CB, Twaroch FA (2010) Multi-source toponym data integration and mediation for a meta-gazetteer service. In: Geographic Information Science, Springer, pp 234–248
[582] Smith PC, Araya-Guerra R, Bublitz C, Parnes B, Dickinson LM, Van Vorst R, Westfall JM, Pace WD (2005) Missing clinical information during primary care visits. Jama 293(5):565–571
[583] Smith TF, Waterman MS (1981) Identification of Common Molecular Subsequences. Molecular Biology 147:195–197

[584] Soto CM, Kleinman KP, Simon SR (2002) Quality and correlates of medical record documentation in the ambulatory care setting. BMC health services research 2(1):22

[585] Soundararajan R, Bovik A (2012) Rred indices: Reduced reference entropic differencing for image quality assessment. IEEE Transactions on Image Processing 21(2):517–526

[586] Soundararajan R, Bovik A (2013) Video quality assessment by reduced reference spatio-temporal entropic differencing. Circuits and Systems for Video Technology, IEEE Transactions on 23(4):684–694

[587] Sriram J, Shin M, Kotz D, Rajan A, Sastry M, Yarvis M (2009) Challenges in data quality assurance in pervasive health monitoring systems. In: Future of Trust in Computing, Springer, pp 129–142

[588] Star SL, Bowker GC (1999) Sorting Things Out: Classification and its Consequences. MIT Press, London, UK

[589] Stelfox HT, Palmisani S, Scurlock C, Orav EJ, Bates DW (2006) The "To err is human" report and the patient safety literature. Quality & safety in health care 15(3):174âĂŞ178, DOI 10.1136/qshc.2006.017947, URL http://www.ncbi.nlm.nih.gov/pubmed/16751466, PMID: 16751466

[590] Stephenson B (1985) Management by information. Information Strategy: The Executive's Journal 1(4):26–32

[591] Stoica M, Chawat N, Shin N (2003) An Investigation of the Methodologies of Business Process Reengineering. In: Proc. of Information Systems Education Conference

[592] Stolfo SJ, Hernandez MA (1995) The Merge/Purge Problem for Large Databases. In: Proc. SIGMOD 1995

[593] Storey V, Wang RY (2001) Extending the ER Model to Represent Data Quality Requirements. In: Wang R, Ziad M, Lee W (eds) Data Quality, Kluver Academic Publishers

[594] Storey VC, Wang RY (1998) An Analysis of Quality Requirements in Database Design. In: Proc. 4th International Conference on Information Quality (IQ 1998)

[595] Strauss A, Fagerhaugh S, Suczek B, Wiener C (1985) The Social Organization of Medical Work. University of Chicago Press, New York, NY, USA

[596] Stvilia B, Mon L, Yi YJ (2009) A model for online consumer health information quality. Journal of the American Society for Information Science and Technology 60(9):1781–1791, DOI 10.1002/asi.21115, URL http://doi.wiley.com/10.1002/asi.21115

[597] Suthaharan S (2009) No-reference visually significant blocking artifact metric for natural scene images. Signal Processing 89(8):1647 – 1652

[598] Talukdar PP, Jacob M, Mehmood MS, Crammer K, Ives ZG, Pereira F, Guha S (2008) Learning to create data-integrating queries. PVLDB 1(1):785–796

[599] Talukdar PP, Ives ZG, Pereira F (2010) Automatically incorporating new sources in keyword search-based data integration. In: SIGMOD Conference 2010, pp 387–398

[600] Tamassia R, Batini C, Di Battista G (1987) Automatic Graph Drawing and Readability of Diagrams. IEEE Transactions on Systems, Men and Cybernetics

[601] Tan WC (2007) Provenance in Databases: Past, Current, and Future. IEEE Data Engineering Bulletin 30(4):3–12

[602] Tarjan RE (1975) Efficiency of A Good But Not Linear Set Union Algorithm. Journal of the ACM 22(2):215–225

[603] TASI (1979) Technical advisory service for images

[604] Tejada S, Knoblock C, Minton S (2001) Learning Object Identication Rules for Information Integration. Information Systems 26(8)

[605] Teperi J (1993) Multi method approach to the assessment of data quality in the finnish medical birth registry. Journal of epidemiology and community health 47(3):242âĂŞ247

[606] Theoharis Y, Fundulaki I, Karvounarakis G, Christophides V (2011) On provenance of queries on semantic web data. IEEE Internet Computing 15(1):31–39

[607] Thiru K, Hassey A, Sullivan F (2003) Systematic review of scope and quality of electronic patient record data in primary care. British Medical Journal 326(7398):1070–1072

[608] Thirunarayan K, Anantharam P, Henson C, Sheth A (2013) Comparative Trust Management with Applications: Bayesian Approaches Emphasis. Future Generation Computer Systems

[609] Thurstone LL (1927) A law of comparative judgement. Psychological Review 34:273–286

[610] Torgerson W (1958) Theory and Methods of Scaling. Wiley, Ney York

[611] Tourancheau S, Autrusseau F, Sazzad Z, Horita Y (2008) Impact of subjective dataset on the performance of image quality metrics. In: Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on, pp 365–368

[612] Trepetin S (2008) Privacy-preserving string comparisons in record linkage systems: a review. Information Security Journal: A Global Perspective 17(5-6):253–266

[613] Ullman JD (1988) Principles of Database and Knowledge-Base Systems. Computer Science Press

[614] Umbrich J, Hausenblas M, Hogan A, Polleres A, Decker S (2010) Towards Dataset Dynamics: Change Frequency of Linked Open Data Sources. In: 3rd Linked Data on the Web Workshop at WWW

[615] UNECE (accessed 2014) `http://www1.unece.org/stat/platform/display/bigdata/Classification+of+Types+of+Big+Data`

[616] Unit EI (2011) Big data: Harnessing a game-changing asset. a report from the economist intelligence unit sponsored by sas

[617] US National Archives (accessed February 09, 2012) Technical guidelines for digitizing archival materials for electronic access: creation of production master files - raster images. URL http://www.archives.gov/preservation/technical/guidelines.html
[618] US National Institute of Health (NIH) (accessed 2014) http://www.pubmedcentral.nih.gov/
[619] Van Engers TM (2004) Legal engineering: A knowledge engineering approach to improving legal quality. eGovernment and eDemocracy: Progress and Challenges pp 189–206
[620] Vantongelen K, Rotmensz N, Van Der Schueren E (1989) Quality control of validity of data collected in clinical trials. European Journal of Cancer and Clinical Oncology 25(8):1241–1247, DOI 10.1016/0277-5379(89)90421-5, URL http://linkinghub.elsevier.com/retrieve/pii/0277537989904215
[621] Vapnik VN, Vapnik V (1998) Statistical learning theory, vol 2. Wiley New York
[622] Vatsalan D, Christen P, Verykios VS (2013) A taxonomy of privacy-preserving record linkage techniques. Information Systems 38(6):946–969
[623] Veregin H, Hargitai P (1995) An evaluation matrix for geographical data quality. Elements of spatial data quality pp 167–188
[624] Verhulst S (2006) Background issues on data quality. Tech. rep., Markle Foundation, URL www.connectingforhealth.org
[625] Verykios VS, Moustakides GV, Elfeky MG (2003) A Bayesian Decision Model for Cost Otimal Record Matching. The VLDB Journal 12:28–40
[626] Verykios VS, Karakasidis A, Mitrogiannis VK (2009) Privacy preserving record linkage approaches. International Journal of Data Mining, Modelling and Management 1(2):206–221
[627] Vessey I (1991) Cognitive fit: A theory-based analysis of the graphs versus tables literature*. Decision Sciences 22(2):219–240
[628] Villata S, Gandon F (2012) Licenses compatibility and composition in the web of data. In: COLD
[629] Vincent C, Neale G, Woloshynowych M (2001) Adverse events in british hospitals: preliminary retrospective record review. Bmj 322(7285):517–519
[630] Viscusi G, Batini C, Mecella M (2010) Information Systems for eGovernment: a Quality of Service Perspective. Springer
[631] VQEG (2000) Final report from the video quality experts group on the validation of objective models of video quality assessment. URL http://www.vqeg.org/
[632] VQEG (2000) Vqeg frtv phase 1 database. URL ftp://ftp.crc.ca/crc/vqeg/TestSequences/
[633] Vydiswaran VGV, Zhai C, Roth D (2011) Content-driven trust propagation framework. In: KDD, pp 974–982

[634] W3C (2013) An overview of the prov family of documents. http://www.w3.org/TR/prov-overview/
[635] W3C (2013) W3c semantic web activity. URL http://www.w3.org/2001/sw/
[636] W3C (accessed 2014) http://www.w3.org/WAI/
[637] Wagner MM, Hogan WR (1996) The accuracy of medication data in an outpatient electronic medical record. Journal of the American Medical Informatics Association 3(3):234–244
[638] Wagner S, Toftegaard TS, Bertelsen OW (2011) Increased data quality in home blood pressure monitoring through context awareness. In: Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2011 5th International Conference on, IEEE, pp 234–237
[639] Wand Y, Wang RY (1996) Anchoring data quality dimensions in ontological foundations. Communications of the ACM 39(11):86–95
[640] Wand Y, Wang RY (1996) Anchoring Data Quality Dimensions in Ontological Foundations. Communications of the ACM 39(11)
[641] Wang J, Kraska T, Franklin MJ, Feng J (2012) Crowder: Crowdsourcing entity resolution. Proceedings of the VLDB Endowment 5(11):1483–1494
[642] Wang RY (1998) A product perspective on total data quality management. Communications of the ACM 41(2):58–65
[643] Wang RY, Madnick SE (1990) A Polygen Model for Heterogeneous Database Systems: The Source Tagging Perspective. In: Proc. VLDB'90, pp 519–538
[644] Wang RY, Strong DM (1996) Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems 12(4)
[645] Wang RY, Strong DM (1996) Beyond accuracy: What data quality means to data consumers. Journal of management information systems pp 5–33
[646] Wang RY, Storey VC, Firth CP (1995) A Framework for Analysis of Data Quality Research. IEEE Transaction on Knowledge and Data Engineering 7(4)
[647] Wang RY, Lee YL, Pipino L, Strong DM (1998) Manage Your Information as a Product. Sloan Management Review 39(4):95–105
[648] Wang RY, Ziad M, Lee YW (2001) Data Quality. Kluwer Academic Publisher
[649] Wang RY, Chettayar K, Dravis F, Funk J, Katz-Haas R, Lee C, Lee Y, Xian X, S B (2005) Exemplifying Business Oppurtunities for Improving Data Quality from Corporate Household Research. In: Wang RY, Pierce EM, Madnick SE, Fisher CW (eds) Advances in Management Information Systems - Information Quality (AMIS-IQ) Monograph, Sharpe, M.E.

[650] Wang RY, Pierce E, Madnick S, Fisher C (2005) Information Quality, Advances in Management Information Systems. M.E. Sharpe, Vladimir Zwass Series
[651] Wang Z, Simoncelli EP (2005) Reduced-reference image quality assessment using a wavelet-domain natural image statistic model. In: in Proc. of SPIE Human Vision and Electronic Imaging, vol 5666, pp 149–159
[652] Wang Z, Bovik A, Evans B (2000) Blind measurement of blocking artifacts in images. In: in Proc. IEEE Int. Conf. Image Proc, pp 981–984
[653] Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image Quality Assessment: From Error Visibility to Structural Similarity. IEEE Transactions on Image Processing 13(4):600–612
[654] Watson AB, Borthwick R, Taylor M (1997) Image quality and entropy masking. In: SPIE Human Vision and Electronic Imaging Conference, vol 3016, pp 2–12
[655] Watts S, Shankaranarayanan G, Even A (2009) Data quality assessment in context: A cognitive perspective. Decision Support Systems 48(1):202–211
[656] Wayne S (1983) Quality control circle and company wide quality control. Quality Progress pp 14–17
[657] Wears RL, Berg M (2005) Computer technology and clinical work: Still waiting for godot. Journal of the American Medical Association 293(10):1261–3
[658] Wee CY, Paramesran R, Mukundan R, Jiang X (2010) Image quality assessment by discrete orthogonal moments. Pattern Recognition 43(12):4055 – 4068
[659] Weis M, Naumann F (2005) DogmatiX Tracks down Duplicates in XML. In: Proc. SIGMOD 2005
[660] Weiskopf NG, Weng C (2013) Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. Journal of the American Medical Informatics Association 20(1):144–151
[661] Whang SE, Garcia-Molina H (2010) Entity resolution with evolving rules. Proceedings of the VLDB Endowment 3(1-2):1326–1337
[662] Whang SE, Garcia-Molina H (2014) Incremental entity resolution on rules and data. The VLDB Journal, The International Journal on Very Large Data Bases 23(1):77–102
[663] Whang SE, Marmaros D, Garcia-Molina H (2013) Pay-as-you-go entity resolution. Knowledge and Data Engineering, IEEE Transactions on 25(5):1111–1124
[664] White C (2005) Data Integration: Using ETL, EAI, and EII Tools to Create an Integrated Enterprise. `http://ibm.ascential.com`
[665] Wiederhold G (1992) Mediators in the Architecture of Future Information Systems. IEEE Computer 25(3)
[666] Wikipedia (accessed 2014) `https://www.wikipedia.org/`

[667] Winkler W (1993) Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage. In: Proc. of the Section on Survey Research Methods, American Statistical Association
[668] Winkler WE (1988) Using the EM Algorithm for Weight Computation in the Fellegi and Sunter Modelo of Record Linkage. In: Proc. of the Section on Survey Research Methods, American Statistical Association
[669] Winkler WE (1995) Matching and record linkage. Business survey methods 1:355–384
[670] Winkler WE (2000) Machine Learning, Information Retrieval and Record Linkage. In: Proc. of the Section on Survey Research Methods, American Statistical Association
[671] Winkler WE (2001) Quality of Very Large Databases. Technical Report RR-2001/04, U.S. Bureau of the Census, Statistical Research Division
[672] Winkler WE (2004) Methods for Evaluating and Creating Data Quality. Information Systems 29(7)
[673] Winkler WE (2006) Overview of record linkage and current research directions. In: Bureau of the Census, Citeseer
[674] Winthereik BR (2003) We fill in our working understanding: on codes, classifications and the production of accurate data. Methods of Information in Medicine 42(4):489–96
[675] Winthereik BR, Vikkels S (2005) ICT and integrated care: some dilemmas of standardising inter-organisational communication. Computer Supported Cooperative Work (CSCW) 14(1):43–67
[676] Wiszniewski B, Krawczyk H (Dublin, Ireland, 2003) Digital Document Life Cycle Development. In: Proc. 1st International Symposium on Information and Communication Technologies (ISICT 2003)
[677] World Health Organization, Regional Office for the Western Pacific (2003) Improving data quality: a guide for developing countries. World Health Organization, Regional Office for the Western Pacific, Manila
[678] Wu W, Yu CT, Doan A, Meng W (2004) An Interactive Clustering-based Approach to Integrating Source Query interfaces on the Deep Web. In: SIGMOD Conference, pp 95–106
[679] Xanthaki H (2001) The problem of quality in eu legislation: what on earth is really wrong? Common Market Law Review 38:651–676
[680] Xue W, Zhang L, Mou X, Bovik AC (2014) Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. Image Processing, IEEE Transactions on 23(2):684–695
[681] Yakout M, Atallah MJ, Elmagarmid A (2009) Efficient private record linkage. In: Data Engineering, 2009. ICDE'09. IEEE 25th International Conference on, IEEE, pp 1283–1286
[682] Yakout M, Elmagarmid AK, Elmeleegy H, Ouzzani M, Qi A (2010) Behavior based record linkage. Proceedings of the VLDB Endowment 3(1-2):439–448
[683] Yan LL, Ozsu T (1999) Conflict Tolerant Queries in AURORA. In: Proc. CoopIS'99

[684] Yan S, Lee D, Kan MY, Giles LC (2007) Adaptive sorted neighborhood methods for efficient record linkage. In: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries, ACM, pp 185–194

[685] Ye P, Doermann D (2012) No-reference image quality assessment using visual codebooks. IEEE Transactions on Image Processing 21(7):3129–3138

[686] Yendrikhovskij S (1999) Image quality: Between science and fiction. In: PICS, pp 173–178

[687] Yin X, Han J (2007) Truth discovery with multiple conflicting information providers on the web. In: In Proc. 2007 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'07)

[688] Zakaluk BL, Samuels SJ (1988) Readability: Its Past, Present, and Future. ERIC

[689] Zardetto D, Scannapieco M, Catarci T (2010) Effective Automated Object Matching. In: Proceedings of the International Conference on Data Engineering (ICDE 2010)

[690] Zardetto D, Valentino L, Scannapieco M (2011) MAERLIN: New Record Linkage Methods At Work. In: Proceedings of the 6th International Conference on New Techniques and Technologies for Statistics (NTTS 2011)

[691] Zaveri A, Rula A, Maurino A, Pietrobon R, Lehmann J, Auer S (2012) Quality assessment methodologies for linked open data (under review). Semantic Web Journal URL `http://www.semantic-web-journal.net/content/quality-assessment-methodologies-linked-open-data`, this article is still under review

[692] Zhang X, Wandell BA (1997) A spatial extension of cielab for digital color-image reproduction. Journal of the Society for Information Display 5(1):61–63

[693] Zhao B, Rubinstein BIP, Gemmell J, Han J (2012) A bayesian approach to discovering truth from conflicting sources for data integration. PVLDB 5(6):550–561

[694] Zhao H, Ram S (2005) Entity identification for heterogeneous database integration—a multiple classifier system approach and empirical evaluation. Information Systems 30(2):119–132

[695] Zingmond DS, Ye Z, Ettner SL, Liu H (2004) Linking hospital discharge and death records—accuracy and sources of bias. Journal of clinical epidemiology 57(1):21–29

Index