Data Wrangling for Big Data: Challenges and Opportunities

Visionary Paper

Data Wrangling for Big Data: Challenges and Opportunities

Tim Furche
Dept. of Computer Science
Oxford University
Oxford OX1 3QD, UK
tim.furche@cs.ox.ac.uk

Georg Gottlob
Dept. of Computer Science
Oxford University
Oxford OX1 3QD, UK
georg.gottlob@cs.ox.ac.uk

Leonid Libkin
School of Informatics
University of Edinburgh
Edinburgh EH8 9AB, UK
libkin@ed.ac.uk

Giorgio Orsi
School. of Computer Science
University of Birmingham
Birmingham, B15 2TT, UK
G.Orsi@cs.bham.ac.uk

Norman W. Paton
School of Computer Science
University of Manchester
Manchester M13 9PL, UK
npaton@manchester.ac.uk

ABSTRACT

Data wrangling is the process by which the data required by an application is identified, extracted, cleaned and integrated, to yield a data set that is suitable for exploration and analysis. Although there are widely used Extract, Transform and Load (ETL) techniques and platforms, they often require manual work from technical and domain experts at different stages of the process. When confronted with the 4 V's of big data (volume, velocity, variety and veracity), manual intervention may make ETL prohibitively expensive. This paper argues that providing cost-effective, highly-automated approaches to data wrangling involves significant research challenges, requiring fundamental changes to established areas such as data extraction, integration and cleaning, and to the ways in which these areas are brought together. Specifically, the paper discusses the importance of comprehensive support for context awareness within data wrangling, and the need for adaptive, pay-as-you-go solutions that automatically tune the wrangling process to the requirements and resources of the specific application.

1. INTRODUCTION

Data wrangling has been recognised as a recurring feature of big data life cycles. Data wrangling has been defined as:

> *a process of iterative data exploration and transformation that enables analysis. ([21])*

In some cases, definitions capture the assumption that there is significant manual effort in the process:

> *the process of manually converting or mapping data from one "raw" form into another format that allows for more convenient consumption of the data with the help of semi-automated tools. ([35])*

The general requirement to reorganise data for analysis is nothing new, with both database vendors and data integration companies providing Extract, Transform and Load (ETL) products [34]. ETL platforms typically provide components for wrapping data sources, transforming and combing data from different sources, and for loading the resulting data into data warehouses, along with some means of orchestrating the components, such as a workflow language. Such platforms are clearly useful, but in being developed principally for enterprise settings, they tend to limit their scope to supporting the specification of wrangling workflows by expert developers.

Does big data make a difference to what is needed for ETL? Although there are many different flavors of big data applications, the 4 V's of big data[1] refer to some recurring characteristics: *Volume* represents scale either in terms of the size or number of data sources; *Velocity* represents either data arrival rates or the rate at which sources or their contents may change; *Variety* captures the diversity of sources of data, including sensors, databases, files and the deep web; and *Veracity* represents the uncertainty that is inevitable in such a complex environment. When all 4 V's are present, the use of ETL processes involving manual intervention at some stage may lead to the sacrifice of one or more of the V's to comply with resource and budget constraints. Currently,

> *data scientists spend from 50 percent to 80 percent of their time collecting and preparing unruly digital data. ([24])*

and only a fraction of an expert's time may be dedicated to value-added exploration and analysis.

In addition to the technical case for research in data wrangling, there is also a significant business case; for example, vendor revenue from big data hardware, software and services was valued at $13B in 2013, with an annual growth rate of 60%. However, just as significant is the nature of the associated activities. The UK Government's Information Economy Strategy states:

> *the overwhelming majority of information economy businesses – 95% of the 120,000 enterprises in the sector – employ fewer than 10 people. ([14])*

As such, many of the organisations that stand to benefit from big data will not be able to devote substantial resources to value-added

[1] `http://www.ibmbigdatahub.com/infographic/four-vs-big-data`.

data analyses unless massive automation of wrangling processes is achieved, e.g., by limiting manual intervention to high-level feedback and to the specification of exceptions.

Example 1 (e-Commerce Price Intelligence)**.** When running an e-Commerce site, it is necessary to understand pricing trends among competitors. This may involve getting to grips with: Volume – thousands of sites; Velocity – sites, site descriptions and contents that are continually changing; Variety – in format, content, targeted community, etc; and Veracity – unavailability, inconsistent descriptions, unavailable offers, etc. Manual data wrangling is likely to be expensive, partial, unreliable and poorly targeted.

As a result, there is a need for research into how to make data wrangling more cost effective. The contribution of this vision paper is to characterise research challenges emerging from data wrangling for the 4Vs (Section 2), to identify what existing work seems to be relevant and where it needs to be further developed (Section 3), and to provide a vision for a new research direction that is a prerequisite for widespread cost-effective exploitation of big data (Section 4).

2. DATA WRANGLING – RESEARCH CHALLENGES

As discussed in the introduction, there is a need for cost-effective data wrangling; the 4 V's of big data are likely to lead to the manual production of a comprehensive data wrangling process being prohibitively expensive for many users. In practice this means that data wrangling for big data involves: (i) *making compromises* – as the perfect solution is not likely to be achievable, it is necessary to understand and capture the priorities of the users and to use these to target resources in a cost-effective manner; (ii) *extending boundaries* – as relevant data may be spread across many organisations and of many types; (iii) *making use of all the available information* – applications differ not only in the nature of the relevant data sources, but also in existing resources that could inform the wrangling process, and full use needs to be made of existing evidence; and (iv) *adopting an incremental, pay-as-you-go approach* – users need to be able to contribute effort to the wrangling process in whatever form they choose and at whatever moment they choose.

The remainder of this section expands on these features, pointing out the challenges that they present to researchers.

2.1 Making Compromises

Faced with an application exhibiting the 4 V's of big data, data scientists may feel overwhelmed by the scale and difficulty of the wrangling task. It will often be impossible to produce a comprehensive solution, so one challenge is to make well informed compromises.

The *user context* of an application specifies functional and non-functional requirements of the users, and the trade-offs between them.

Example 2 (e-Commerce User Contexts)**.** In price intelligence, following on from Example 1, there may be different user contexts. For example, routine *price comparison* may be able to work with a subset of high quality sources, and thus the user may prefer features such as *accuracy* and *timeliness* to *completeness*. In contrast, where sales of a popular item have been falling, the associated *issue investigation* may require a more complete picture for the product in question, at the risk of presenting the user with more incorrect or out-of-date data.

Thus a single application may have different *user contexts*, and any approach to data wrangling that hard-wires a process for selecting and integrating data risks the production of data sets that are not always fit for purpose. Making well informed compromises involves: (i) capturing and making explicit the requirements and priorities of users; and (ii) enabling these requirements to permeate the wrangling process. There has been significant work on decision-support, for example in relation to multi-criteria decision making [37], that provides both languages for capturing requirements and algorithms for exploring the space of possible solutions in ways that take the requirements into account. For example, in the widely used Analytic Hierarchy Process [31], users compare criteria (such as timeliness or completeness) in terms of their relative importance, which can be taken into account when making decisions (such as which mappings to use in data integration).

Although data management researchers have investigated techniques that apply specific user criteria to inform decisions (e.g. for selecting sources based on their anticipated financial value [16]) and have sometimes traded off alternative objectives (e.g. precision and recall for mapping selection and refinement [5]), such results have tended to address specific steps within wrangling in isolation, often leading to bespoke solutions. Together with high automation, adaptivity and multi-criteria optimisation are of paramount importance for cost-effective wrangling processes.

2.2 Extending the Boundaries

ETL processes traditionally operate on data lying within the boundaries of an organisation or across a network of partners. As soon as companies started to leverage big data and data science, it became clear that data outside the boundaries of the organisation represent both new business opportunities as well as a means to optimize existing business processes.

Data wrangling solutions recently started to offer connectors to external data sources but, for now, mostly limited to open government data and established social networks (e.g., Twitter) via formalised APIs. This makes wrangling processes dependent on the availability of APIs from third parties, thus limiting the availability of data and the scope of the wrangling processes.

Recent advances in web data extraction [19, 30] have shown that fully-automated, large scale collection of long-tail, business-related data, e.g., products, jobs or locations, is possible. The challenge for data wrangling processes is now to make proper use of this wealth of "wild" data by coordinating extraction, integration and cleaning processes.

Example 3 (Business Locations)**.** Many social networks offer the ability for users to check-in to places, e.g., restaurants, offices, cinemas, via their mobile apps. This gives to social networks the ability to maintain a database of businesses, their locations, and profiles of users interacting with them that is immensely valuable for advertising purposes. On the other hand, this way of acquiring data is prone to data quality problems, e.g., wrong geo-locations, misspelled or fantasy places. A popular way to address these problems is to acquire a curated database of geo-located business locations. This is usually expensive and does not always guarantee that the data is really clean, as its quality depends on the quality of the (usually unknown) data acquisition and curation process. Another way is to define a wrangling process that collects this information right on the website of the business of interest, e.g., by wrapping the target data source directly. The extraction process can in this case be "informed" by existing integrated data, e.g., the business url and a database of already known addresses, to identify previously unknown locations and correct erroneous ones.

2.3 Using All the Available Information

Cost-effective data wrangling will need to make extensive use of

automation for the different steps in the wrangling process. Automated processes must take advantage of all available information both when generating proposals and for comparing alternative proposals in the light of the user context.

The *data context* of an application consists of the sources that may provide data for wrangling, and other information that may inform the wrangling process.

Example 4 (e-Commerce Data Context)**.** In price intelligence, following on from Example 1, the data context includes the catalogs of the many online retailers that sell overlapping sets of products to overlapping markets. However, there are additional data resources that can inform the process. For example, the e-Commerce company has a product catalog that can be considered as master data by the wrangling process; the company is interested in price comparison only for the products it sells. In addition, for this domain there are standard formats, for example in `schema.org`, for describing products and offers, and there are ontologies that describe products, such as The Product Types Ontology[2].

Thus applications have different *data contexts*, which include not only the data that the application seeks to use, but also local and third party sources that provide additional information about the domain or the data therein. To be cost-effective, automated techniques must be able to bring together all the available information. For example, a product types ontology could be used to inform the selection of sources based on their relevance, as an input to the matching of sources that supplements syntactic matching, and as a guide to the fusion of property values from records that have been obtained from different sources. To do this, automated processes must make well founded decisions, integrating evidence of different types. In data management, there are results of relevance to data wrangling that assimilate evidence to reach decisions (e.g. [36]), but work to date tends to be focused on small numbers of types of evidence, and individual data management tasks. Cost effective data wrangling requires more pervasive approaches.

2.4 Adopting a Pay-as-you-go Approach

As discussed in Section 1, potential users of big data will not always have access to substantial budgets or teams of skilled data scientists to support manual data wrangling. As such, rather than depending upon a continuous labor-intensive wrangling effort, to enable resources to be deployed on data wrangling in a targeted and flexible way, we propose an incremental, pay-as-you-go approach, in which the "payment" can take different forms.

Providing a pay-as-you-go approach, with flexible kinds of payment, means automating all steps in the wrangling process, and allowing feedback in whatever form the user chooses. This requires a flexible architecture in which feedback is combined with other sources of evidence (see Section 2.3) to enable the best possible decisions to be made. Feedback of one type should be able to inform many different steps in the wrangling process – for example, the identification of several correct (or incorrect) results may inform both source selection and mapping generation. Although there has been significant work on incremental, pay-as-you-go approaches to data management, building on the dataspaces vision [18], typically this has used one or a few types of feedback to inform a single activity. As such, there is significant work to be done to provide a more integrated approach in which feedback can inform all steps of the wrangling process.

Example 5 (e-Commerce Pay-as-you-go)**.** In Example 1, automated approaches to data wrangling can be used to select sources of product data, and to fuse the values from such sources to provide reports on the pricing of different products. These reports are studied by the data scientists of the e-Commerce company who are reviewing the pricing of competitors, who can annotate the data values in the report, for example, to identify which are correct or incorrect, along with their relevance to decision-making. Such feedback can trigger the data wrangling system to revise the way in which such reports are produced, for example by prioritising results from different data sources. The provision of domain-expert feedback from the data scientists is a form of payment, as staff effort is required to provide it. However, it should also be possible to use crowdsourcing, with direct financial payment of crowd workers, for example to identify duplicates, and thereby to refine the automatically generated rules that determine when two records represent the same real-world object [20]. It is of paramount importance that these feedback-induced "reactions" do not trigger a re-processing of all datasets involved in the computation but rather limit the processing to the strictly necessary data.

[2] `http://www.productontology.org/`

3. DATA WRANGLING – RELATED WORK

As discussed in Section 2, cost-effective data wrangling is expected to involve best-effort approaches, in which multiple sources of evidence are combined by automated techniques, the results of which can be refined following a pay-as-you-go approach. Space precludes a comprehensive review of potentially relevant results, so in this section we focus on three areas with overlapping requirements and approaches, pointing out existing results on which data wrangling can build, but also areas in which these results need to be extended.

3.1 Knowledge Base Construction

In *knowledge base construction* (KBC) the objective is to automatically create structured representations of data, typically using the web as a source of facts for inclusion in the knowledge base. Prominent examples include YAGO [33], Elementary [28] and Google's Knowledge Vault [15], all of which combine candidate facts from web data sources to create or extend descriptions of entities. Such proposals are relevant to data wrangling, in providing large scale, automatically generated representations of structured data extracted from diverse sources, taking account of the associated uncertainties.

These techniques have produced impressive results but they tend to have a single, implicit *user context*, with a focus on consolidating slowly-changing, common sense knowledge that leans heavily on the assumption that correct facts occur frequently (instance-based redundancy). For data wrangling, the need to support diverse user contexts and highly transient information (e.g., pricing) means that user requirements need to be made explicit and to inform decision-making throughout automated processes. In addition, the focus on fully automated KBC at web-scale, without systematic support for incremental improvement in a pay-as-you-go manner, tends to require expert input, for example through the writing of rules (e.g., [28]). As such, KBC proposals share requirements with data wrangling, but have different emphases.

3.2 Pay-as-you-go Data Management

Pay-as-you-go data management, as represented by the dataspaces vision [18], involves the combination of an automated *bootstrapping* phase, followed by incremental *improvement*. There have been numerous results on different aspects of pay-as-you-go data management, across several activities of relevance to data wran-

gling, such as data extraction (e.g., [12]), matching [26], mapping [5] and entity resolution [20]. We note that in these proposals a single type of feedback is used to support a single data management task. The opportunities presented by crowdsourcing have provided a recent boost to this area, in which, typically, paid micro-tasks are submitted to public crowds as a source of feedback for pay-as-you-go activities. This has included work that refines different steps within an activity (e.g. both blocking and fine-grained matching within entity resolution [20]), and the investigation of systematic approaches for relating uncertain feedback to other sources of evidence (e.g., [13]). However, the state-of-the-art is that techniques have been developed in which individual types of feedback are used to influence specific data management tasks, and there seems to be significant scope for feedback to be integrated into all activities that compose a data wrangling pipeline, with reuse of feedback to inform multiple activities [6]. Highly automated wrangling processes require formalised feedback (e.g., in terms of rules or facts to be added/removed from the process) so that they can be used by suitable reasoning processes to automatically adapt the wrangling workflows.

Data Tamer [32] provides a substantially automated pipeline involving schema integration and entity resolution, where components obtain feedback to refine the results of automated analyses. Although Data Tamer moves a significant way from classical, largely manually specified ETL techniques, user feedback is obtained for and applied to specific steps (and not shared across components), and there is no user context to inform where compromises should be made and efforts focused.

3.3 Context Awareness

There has been significant prior work on context in computing systems [3], with a particular emphasis on mobile devices and users, in which the objective is to provide data [9] or services [25] that meet the evolving, situational needs of users. In information management, the emphasis has been on identifying the portion of the available information that is relevant in specific ambient conditions [8]. For data wrangling, classical notions of context such as location and time will sometimes be relevant, but we anticipate that for data wrangling: (i) there may be many additional features that characterise the *user and data contexts*, for individual users, groups of users and tasks; and (ii) that the information about context will need to inform a wide range of data management tasks in addition to the selection of the most relevant results.

4. DATA WRANGLING – VISION

In the light of the scene-setting from the previous sections, Figure 1 outlines potential components and relationships in a data wrangling architecture. To the left of the figure, several (potentially many) *Data Source*s provide the data that is required for the application. A *Data Extraction* component provides wrappers for the potentially heterogeneous sources (files, databases, documents, web pages), providing syntactically consistent representations that can then be brought together by the *Data Integration* component, to yield *Wrangled Data* that is then available for exploration and analysis.

However, in our vision, these extraction and integration components both *use all the available data* and *adopt a pay-as-you-go approach*. In Figure 1, this is represented by a collection of *Working Data*, which contains not only results and metadata from the *Data Extraction* and *Data Integration* components, but also:

1. all relevant *Auxiliary Data*, which would include the *user context*, and whatever additional information can represent the *data context*, such as *reference data*, *master data* and *domain ontologies*;
2. the results of all *Quality* analyses that have been carried out, which may apply to individual data sources, the results of different extractions and components of relevance to integration such as matches or mappings; and
3. the feedback that has been obtained from users or crowds, on any aspect of the wrangling process, including the extractions (e.g. users could indicate if a wrapper has extracted what they would have expected), or the results of integrations (e.g. crowds could identify duplicates).

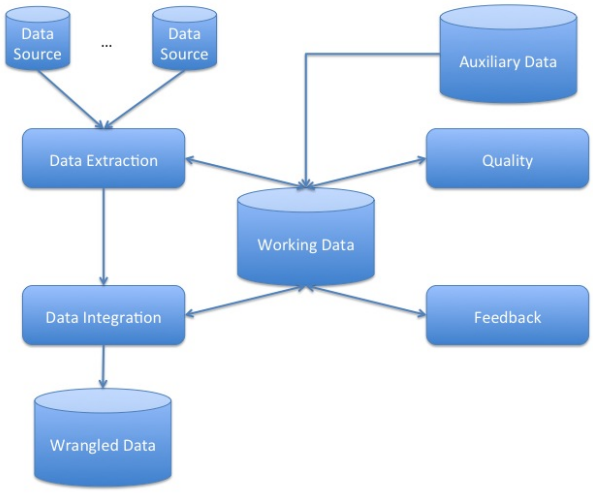

Figure 1: Abstract Wrangling Architecture.

To support this, data wrangling needs substantial advances in data extraction, integration and cleaning, as well as the co-design of the components in Figure 1 to support a much closer interaction in a context-aware, pay-as-you-go setting.

4.1 Research Challenges for Components

This section makes the case that meeting the vision will require changes of substance to existing data management functionalities, such as *Data Extraction* and *Data Integration*.

To respond fully to the proposed architecture, *Data Extraction* must make effective use of all the available data. Consider web data extraction, in which wrappers are generated that enable deep web resources to be treated as structured data sets (e.g., [12, 19]). The lack of context and incrementality in data extraction has long been identified as a weakness [11], and research is required to make extraction components responsive to quality analyses, insights from integration and user feedback. As an example, existing knowledge bases and intermediate products of data cleaning and integration processes can be used to improve the quality of wrapper induction (e.g. [29]).

Along the same lines, *Data Integration* must make effective use of all the available data in ways that take account of the *user context*. As data integration acts on a variety of constructs (sources, matches, mappings, instances), each of which may be associated with its own uncertainties, automated functionalities such as those for identifying matches and generating mappings need to be revised to support multi-criteria decision making in the context of uncertainty. For example, the selection of which mappings to use must take into account information from the user context, such as the number of results required, the budget for accessing sources, and quality requirements. To support the complete data wrangling

process involves generalising from a range of point solutions into an approach in which all components can take account of a range of different sources of evolving evidence.

4.2 Research Challenges for Architectures

This section makes the case that meeting the vision will require changes of substance to existing data management architectures, and in particular a paradigm-shift for ETL.

Traditional ETL operates on manually-specified data manipulation workflows that extract data from structured data sources, integrating, cleaning, and eventually storing them in aggregated form into data warehouses. In Figure 1 there is no explicit control flow specified, but we note that the requirements of automation, refined on a pay-as-you-go basis taking into account the user context, is at odds with a hard-wired, user-specified data manipulation workflow. In the abstract architecture, the pay-as-you-go approach is achieved by storing intermediate results of the ETL process for on-demand recombination, depending on the user context and the potentially continually evolving data context. As such, the *user context* must provide a declarative specification of the user's requirements and priorities, both functional (data) and non-functional (such as quality and cost trade-offs), so that the components in Figure 1 can be automatically and flexibly composed. Such an approach requires an autonomic approach to data wrangling, in which self-configuration is more central to the architecture than in self-managing databases [10].

The resulting architecture must not only be autonomic, it must also take account of the inherent uncertainty associated with much of the *Working Data* in Figure 1. Uncertainty comes from: (i) *Data Sources* in the form of unreliable and inconsistent data; (ii) the wrangling components, for example in the form of tentative extraction rules or mappings; (iii) the auxiliary data, for example in the form of ontologies that do not quite represent the user's conceptualisation of the domain; and (iv) the feedback which may be unreliable or out of line with the user's requirements or preferences. With this complex environment, it is important that uncertainty is represented explicitly and reasoned with systematically, so that well informed decisions can build on a sound understanding of the available evidence.

This raises an additional research question, on how best to represent and reason in a principled and scalable way with the *working data* and associated *workflows*; there is a need for a uniform representation for the results of the different components in Figure 1, which are as diverse as domain ontologies, matches, data extraction and transformation rules, schema mappings, user feedback and provenance information, along with their associated quality annotations and uncertainties.

In addition, the ways in which different types of user engage with the wrangling process is also worthy of further research. In Wrangler [22], now commercialised by Trifacta, data scientists clean and transform data sets using an interactive interface in which, among other things, the system can suggest generic transformations from user edits. In this approach, users provide feedback on the changes to the selected data they would like to have made, and select from proposed transformations. Additional research could investigate where such interactions could be used to inform upstream aspects of the wrangling process, such as source selection or mapping generation, and to understand how other kinds of feedback, or the results of other analyses, could inform what is offered to the user in tools such as Wrangler.

4.3 Research Challenges in Scalability

In this paper we have proposed responding to the *Volume* aspect of big data principally in the form of the number of sources that may be available, where we propose that automation and incrementality are key approaches. In this section we discuss some additional challenges in data wrangling that result from scale.

The most direct impact of scale in big data results from the sheer volume of data that may be present in the sources. ETL vendors have responded to this challenge by compiling ETL workflows into big data platforms, such as map/reduce. In the architecture of Figure 1, it will be necessary for extraction, integration and data querying tasks to be able to be executed using such platforms. However, there are also fundamental problems to be addressed. For example, many quality analyses are intractable (e.g. [7]), and evaluating even standard queries of the sort used in mappings may require substantial changes to classical assumptions when faced with huge data sets. Among these challenges are understanding the requirement for query scalability [2] that can be provided in terms of access and indexing information [17], and developing static techniques for query approximation (i.e., without looking at the data) as was initiated in [4] for conjunctive queries. For the architecture of Figure 1 there is the additional requirement to reason with uncertainty over potentially numerous sources of evidence; this is a serious issue since even in the classical settings data uncertainty often leads to intractability of the most basic data processing tasks [1, 23]. We also observe that knowledge base construction has itself given rise to novel reasoning techniques [27], and additional research may be required to inform decision-making for data wrangling at scale.

5. CONCLUSIONS

Data wrangling is a problem and an opportunity:

- A problem because the 4 V's of big data may all be present together, undermining manual approaches to ETL.
- An opportunity because if we can make data wrangling much more cost effective, all sorts of hitherto impractical tasks come into reach.

This vision paper aims to raise the profile of data wrangling as a research area within the data management community, where there is a lot of work on relevant functionalities, but where these have not been refined or integrated as is required to support data wrangling. The paper has identified research challenges that emerge from data wrangling, around the need to make compromises that reflect the user's requirements, the ability to make use of all the available evidence, and the development of pay-as-you-go techniques that enable diverse forms of payment at convenient times. We have also presented an abstract architecture for data wrangling, and outlined how that architecture departs from traditional approaches to ETL, through increased use of automation, which flexibly accounts for diverse user and data contexts. It has been suggested that this architecture will require changes of substance to established data management components, as well as the way in which they work together. For example, the proposed architecture will require support for representing and reasoning with the diverse and uncertain working data that is of relevance to the data wrangling process. Thus we encourage the data management research community to direct its attention at novel approaches to data wrangling, as a prerequisite for the cost-effective exploitation of big data.

Acknowledgments

This research is supported by the VADA Programme Grant from the UK Engineering and Physical Sciences Research Council, whose support we are pleased to acknowledge. We are also grateful to our colleagues in VADA for their contributions to discussions on data wrangling: Peter Buneman, Wenfei Fan, Alvaro Fernandes, John

Keane, Thomas Lukasiewicz, Sebastian Maneth and Dan Olteanu.

6. REFERENCES

[1] S. Abiteboul, P. Kanellakis, and G. Grahne. On the representation and querying of sets of possible worlds. *TCS*, 78(1):158–187, 1991.

[2] M. Armbrust, E. Liang, T. Kraska, A. Fox, M. J. Franklin, and D. A. Patterson. Generalized scale independence through incremental precomputation. In *SIGMOD*, pages 625–636, 2013.

[3] M. Baldauf, S. Dustdar, and F. Rosenberg. A survey on context-aware systems. *IJAHUC*, 2(4):263–277, 2007.

[4] P. Barceló, L. Libkin, and M. Romero. Efficient approximations of conjunctive queries. *SIAM J. Comput.*, 43(3):1085–1130, 2014.

[5] K. Belhajjame, N. W. Paton, S. M. Embury, A. A. A. Fernandes, and C. Hedeler. Incrementally improving dataspaces based on user feedback. *Inf. Syst.*, 38(5):656–687, 2013.

[6] K. Belhajjame, N. W. Paton, A. A. A. Fernandes, C. Hedeler, and S. M. Embury. User feedback as a first class citizen in information integration systems. In *CIDR*, pages 175–183, 2011.

[7] P. Bohannon, W. Fan, M. Flaster, and R. Rastogi. A cost-based model and effective heuristic for repairing constraints by value modification. In *SIGMOD*, pages 143–154, 2005.

[8] C. Bolchini, C. Curino, E. Quintarelli, F. A. Schreiber, and L. Tanca. A data-oriented survey of context models. *SIGMOD Rec.*, 36(4):19–26, 2007.

[9] C. Bolchini, C. A. Curino, G. Orsi, E. Quintarelli, R. Rossato, F. A. Schreiber, and L. Tanca. And what can context do for data? *CACM*, 52(11):136–140, 2009.

[10] S. Chaudhuri and V. R. Narasayya. Self-tuning database systems: A decade of progress. In *VLDB*, pages 3–14, 2007.

[11] S. Chuang, K. C. Chang, and C. X. Zhai. Collaborative wrapping: A turbo framework for web data extraction. In *ICDE*, 2007.

[12] V. Crescenzi, P. Merialdo, and D. Qiu. A framework for learning web wrappers from the crowd. In *WWW*, pages 261–272, 2013.

[13] G. Demartini, D. E. Difallah, and P. Cudré-Mauroux. Large-scale linked data integration using probabilistic reasoning and crowdsourcing. *VLDBJ*, 22(5):665–687, 2013.

[14] Department for Business, Innovation & Skills. Information economy strategy. `http://bit.ly/1W4TPGU`, 2013.

[15] X. L. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, , and W. Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *KDD*, pages 601–610, 2014.

[16] X. L. Dong, B. Saha, and D. Srivastava. Less is more: Selecting sources wisely for integration. *PVLDB*, 6(2):37–48, 2012.

[17] W. Fan, F. Geerts, and L. Libkin. On scale independence for querying big data. In *PODS*, pages 51–62, 2014.

[18] M. J. Franklin, A. Y. Halevy, and D. Maier. From databases to dataspaces: a new abstraction for information management. *SIGMOD Record*, 34(4):27–33, 2005.

[19] T. Furche, G. Gottlob, G. Grasso, X. Guo, G. Orsi, C. Schallhart, and C. Wang. DIADEM: Thousands of websites to a single database. *PVLDB*, 7(14):1845–1856, 2014.

[20] C. Gokhale, S. Das, A. Doan, J. F. Naughton, N. Rampalli, J. W. Shavlik, and X. Zhu. Corleone: hands-off crowdsourcing for entity matching. In *SIGMOD*, pages 601–612, 2014.

[21] S. Kandel, J. Heer, C. Plaisant, J. Kennedy, F. van Ham, N. H. Riche, C. Weaver, B. Lee, D. Brodbeck, and P. Buono. Research directions in data wrangling: Visuatizations and transformations for usable and credible data. *Information Visualization*, 10(4):271–288, 2011.

[22] S. Kandel, A. Paepcke, J. Hellerstein, and J. Heer. Wrangler: Interactive visual specification of data transformation scripts. In *CHI*, pages 3363–3372, 2011.

[23] L. Libkin. Incomplete data: what went wrong, and how to fix it. In *PODS*, pages 1–13, 2014.

[24] S. Lohr. For big-data scientists, 'janitor work' is key hurdle to insights. `http://nyti.ms/1Aqif2X`, 2014.

[25] Z. Maamar, D. Benslimane, and N. C. Narendra. What can context do for web services? *CACM*, 49(12):98–103, 2006.

[26] R. McCann, W. Shen, and A. Doan. Matching schemas in online communities: A web 2.0 approach. In *ICDE*, pages 110–119, 2008.

[27] F. Niu, C. Ré, A. Doan, and J. W. Shavlik. Tuffy: Scaling up statistical inference in markov logic networks using an RDBMS. *PVLDB*, 4(6):373–384, 2011.

[28] F. Niu, C. Zhang, C. Ré, and J. W. Shavlik. Elementary: Large-scale knowledge-base construction via machine learning and statistical inference. *IJSWIS*, 8(3):42–73, 2012.

[29] S. Ortona, G. Orsi, M. Buoncristiano, and T. Furche. Wadar: Joint wrapper and data repair. *PVLDB*, 8(12):1996–2007, 2015.

[30] D. Qiu, L. Barbosa, X. L. Dong, Y. Shen, and D. Srivastava. DEXTER: large-scale discovery and extraction of product specifications on the web. *PVLDB*, 8(13):2194–2205, 2015.

[31] T. L. Saaty. The modern science of multicriteria decision making and its practical applications: The AHP/ANP approach. *Operations Research*, 61(5):1101–1118, 2013.

[32] M. Stonebraker, D. Bruckner, I. F. Ilyas, G. Beskales, M. Cherniack, S. B. Zdonik, A. Pagan, and S. Xu. Data curation at scale: The data tamer system. In *CIDR 2013, Sixth Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 6-9, 2013, Online Proceedings*, 2013.

[33] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge. In *WWW*, pages 697–706, 2007.

[34] P. Vassiliadis. A survey of extract-transform-load technology. *IJDWM*, 5(3):1–27, 2011.

[35] Wikipedia: Various Authors. Data wrangling. `http://bit.ly/1KslZb7`, 2007.

[36] X. Yin, J. Han, and P. S. Yu. Truth discovery with multiple conflicting information providers on the web. *IEEE Trans. Knowl. Data Eng.*, 20(6):796–808, 2008.

[37] C. Zopounidis and P. M. Pardalos. *Handbook of Multicriteria Analysis*. Springer, 2010.