Survey: When semantics meet Crowdsourcing to enhance Big Data Variety

1st Houssein Dhayne
CIMTI, ESIB
Saint-Joseph University
Beirut, Lebanon
houssein.dhayne@net.usj.edu.lb

2nd Rima Kilany Chamoun
CIMTI, ESIB
Saint-Joseph University
Beirut, Lebanon
rima.kilany@usj.edu.lb

3rd Maria Sokhn
Valais Wallis, Technopole 3
Haute École spécialisée de Suisse occidentale
Sierre, Switzerland
maria.sokhn@hes-so.ch

***Abstract*—With the rapid growth of collected data and the variety of its content, the need for efficient integration at a Big Data level becomes crucial. Semantic technologies, as a means of integration and coordination of heterogeneous systems, may help big data to manage terminology and relationships to link various data from different data sources. However, and due to the difficulty of integration and analytics of some datasets with high-precision, automated processes cannot reach a high level of accuracy without the human cognitive ability. Crowdsourcing platforms have the potential to integrate (entity matching, entity resolution) and analyze (sentiment analysis, image recognition) heterogeneous data sources when in some cases these integration tasks may prove to be problematic for computers. In this survey, we explore and compare empirical research studies that rely on merging semantic and crowdsourcing technologies. And, in the light of this comparison, we propose a high-level integration workflow, which shows how merging these technologies can enhance the big data integration process and tackle the data analysis challenges.**

***Index Terms*—Big Data Integration, Semantic Web, Crowdsourcing, Linked Open Data**

I. Introduction

Big Data is characterized by three basic properties (3 Vs.): Volume, Velocity and Variety [1] [2]. Some additional features like Variability and Complexity [3], as well as Value and Veracity [4] are also associated to the concept of Big Data. Another element to take into consideration is the fact that the value of data increases exponentially when it is linked and fused with other data. Hence addressing the data integration challenge is critical to realizing the promise of Big Data [5], and unfortunately existing data warehousing techniques are inefficient to handle such integration [6]. Indeed, traditional data warehouses integrate structured, transactional data that is contained within relational databases. In contrast, unstructured data, which comes in the form of emails, social media, blogs, documents, images, and videos need a novel methodology for the integration process. In recent years, a new concept of data lakes has appeared that stores a vast amount of raw data in its native format and support flexible "schema-on-read" with the help of metadata descriptions [7]. Nevertheless, preparing, organizing, exploring, and querying a data lake is often strenuous. In particular, "on-demand integration" does not take into account the need for data quality or schema understanding [8].

To meet this gap, semantic technologies are seen as an ideal solution to big data integration. In [9], Shadbolt explains that the major driver for the semantic web has been the integration between diverse and heterogeneous data sets that come from different scientific communities; this is being achieved through the adoption of common conceptualizations referred to as ontologies. Unfortunately, data-accuracy problems affect those datasets, caused by incomplete or incorrect data, inconsistencies, etc. These problems arise during the creation process of semantic data, due to errors in the original data source, the tools employed to convert or create semantic data, and the misuse of ontologies, etc. [10].

Crowdsourcing has lately emerged as a viable platform for data integration techniques [11], it is an effective tool to help produce such high-quality data [12]. Indeed, it benefits from the intelligence of online communities to solve a specific problem or complete a task [13]. D. Brabham in [14] identifies four specific approaches of crowdsourcing:

- Knowledge discovery and management: Finding and collecting information through an on-line community into a common location and format.
- Broadcast search: It involves broadcasting a problem-solving challenge widely on the internet and offering an award for the solution.
- Peer-vetted creative production: The on-line community both proposes possible solutions and is empowered to vote on the ideas in order to collectively choose among the solutions.
- Distributed human intelligence tasking: Analyzing large amounts of information, where human intelligence/computing, is more efficient or effective than computer analysis.

The rest of the paper discusses the benefits of associating a semantic approach along with crowdsourcing in order to improve data integration despite its variety. In section II we describe the benefits of such a combination. In section III we detail some use cases that can be described as success stories for such an association. In section IV we analyze and compare the methods and techniques used for data integration in the use cases we exposed in the previous section. In section V we propose a generic workflow for data integration based

on merging these crowdsourcing and linked data techniques. Finally section VI concludes this paper and presents a research proposal that would benefit from the hereby described combination.

II. SEMANTIC APPRAOCH MEETS CROWDSOURCING

Current research clearly indicates that crowdsourcing and the Semantic web are complementary to one another. According to [15] crowdsourcing can be a valuable tool in the Semantic Web endeavor:

- It can improve accuracy of existing automatic techniques by offering a systematic way to augment these techniques with human inputs. RDF-Hunter [16] is an example that implements query decomposition techniques to automatically decide the parts of a query that should resort to the crowd.
- It can help achieve complex and long tasks, in a scalable and affordable way, by distributing tasks to a large number of contributors and using novel incentive models to encourage participation.
- It allows exploiting the cognitive diversity of collective intelligence.

According to [17], The Global Brain Semantic Web - a Semantic web interleaving a large number of human and machine computations - has great potential to overcome some of the issues of the current Semantic Web. In [18], the authors assert that human computation can be used to help curate the semantic web of data in the Linked Open Data (LOD) cloud, so the semantic web can be used to provide better user continuity and platform consistency across human computation systems. In the opposite direction, the Semantic Web can mainly offer three core contributions to crowdsourcing tools [15]:

- Machine-processable semantics facilitate the formal explicit specification of the crowdsourcing domain with all its components.
- Linked Data standards and protocols facilitate information integration and reuse across crowdsourcing platforms and experiments.
- Reasoning could enhance the capabilities of specific crowdsourcing-related methods.

III. BIG DATA INTEGRATION USE CASES WHERE SEMANTICS MEET CROWDSOURCING

In the previous section we showed the general benefits that could be achieved by the combination of semantics and crowdsourcing. In this section, we detail some specific use cases (cf. Table I) of research developments that try to overcome Big Data integration and analysis challenges by taking advantage of this combination.

The authors in [19] point out to the fact that *geosciences* entered the realm of big data due to the huge amount of research data being shared in open repositories (i.e. data, presentations, code). They have already amassed over 30 million semantic statements into Linked Open Data datasets, describing conference attendees, co-authorship, professional society membership, meetings attended, and other related information regarding the geoscience research network. But although semantics can enable semi-automated alignment between data repository and provide means to link all this data, nevertheless for some tasks, semantic algorithms do not reach the needed level of accuracy, and this is where augmenting these algorithms with crowdsourcing can be useful as the authors argue. Therefore, they have developed a crowdsourcing portal that allows members of the geoscience community to link their conference information and grant descriptions, to available LOD datasets. The user input is converted into RDF, and these links are then deployed in subsequent data discovery tools.The crowd, in this case, is comprised of professional researchers and not the general public. This fact led the authors to tackle a key challenge to crowdsourcing: incentives. The authors propose the use of Altmetrics [20] which is an attempt to extend a researchers profile by quantifying scholarly impact beyond the traditional journal citation. The second challenge that was studied is the possibility to evaluate the quality of the crowd sourced data, which can vary considerably. Automated metrics such as Fleisskappa [21] in conjunction with semantic provenance, for assessing the value of the crowd would be a solution. The last challenge to be studied by the authors is Annotation, Trust and Provenance. This is essentially about how data of crowdsourcing is aligned with existing knowledge base triples, and how to handle multiple annotations of the same knowledge base triple. The semantics of trust and provenance need to be captured within semantic crowdsourcing applications. Trust cannot replace provenance, and vice versa.

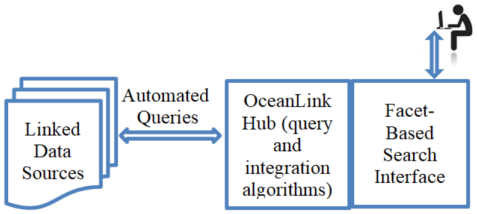

Fig. 1. Oceanlink infrastructure

The OceanLink project [22] is in fact a particular geoscience project. It tackles the challenge of integrating the diverse, complex and huge number of scholarly products and achievements produced by ocean science investigations -such as published articles, datasets, software, and associated supporting materials- by leveraging Semantic Web technologies, web mining, and crowdsourcing to identify links between data centers, digital repositories, and professional societies. Enhancing discovery, enabling collaboration, and assessing research contribution are all addressed by an online platform that applies semantic technologies to support data representation, discovery, sharing and integration. Furthermore, and in order to provide large scale horizontal integration, the OceanLink project uses linked data and Ontology Design Patterns (ODPs). The ODP approach advocates a set of partial ontologies, each

of which formalizes only one key notion that can be aligned with the different representation choices that has been made in different repositories. To meet the challenge of querying multiple LOD sources, the OceanLink project builds a hub to automatically identify links between datasets, to ensure co-entity resolution, and to provide a user-friendly facet-based search interface (see fig. 1). Another benefit of inference/reasoning resulting from the use of LOD and ODP will be for quality assurance, since semantic tools can reason/infer that data is inconsistent. Data may not only be inconsistent, but also messy and not all of the inferred links between entities may be valid. This is where the users are solicited to help validate links, and this is where in this case crowdsourcing would benefit semantic reasoning. OceanLink is currently scaling out to tens of data providers and an estimated one billion semantic statements. In conclusion, this project is providing a building block for future semantic geoscience integration with a crowdsourcing support. It is intended to be applicable beyond the ocean sciences and to be used in other Big Data scenarios

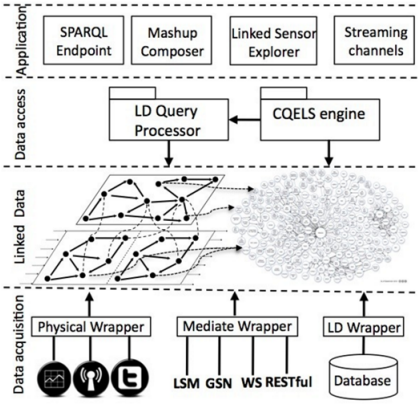

Fig. 2. Linked Sensor Middleware architecture [23].

An *urban environments* project [23] offers to integrate the huge amount of User Generated Content (UGC) and location of the users available through the use of social media and smartphones, in order to offer new spatial business intelligence decision tools for various domain-specific applications. Authors describe the Linked Sensor Framework (LSF) a framework to aggregate and link heterogeneous geolocated data from various sources by using, Semantic Web technologies and transformation to a Linked Data format. Possible applications of LSF may include decision support systems for emergency response, smart cities and tourism or spatial crowdsourcing through integration of UGC (opinions, social media data, surveys) with sensor data and other sources, allowing for example, collection of opinions, sentiment and ideas from particular area of London about best dates for closure of a given subway station for maintenance, or changes in public transport timetables (see fig.2). Other applications include mobile social reporting, where citizens can report issues within their locality or issues they experience while traveling i.e. in their local commercial area.

Urban-related information and geographic data such as interesting facets of cities, street topology, road traffic conditions, business activities, points of interest, etc, are more and more studied in the Linked Data community. In paper [24], the authors introduce the UrbanMatch mobile location-aware game with a Purpose, which tries to alleviate main drawbacks hampering a larger adoption of Linked Data in Smart Cities scenarios. The two main drawbacks are according to the authors: the doubtful quality of the available information thus making people distrust Linked Data content, and the lack of user-centered tools preventing people from contributing. UrbanMatch engages players to provide information specific to the city of Milano. It is aimed at linking points of interests in the city with the most representative photos retrieved from social media Web sites and to rank those links, so to identify the most characteristic ones and to discard the others, thus improving the quality. The input data come from available Web sources (cf. fig.3). Points of interest in Milano were collected and chosen among those available from OpenStreetMap; an RDF description of those POIs is also available in Linked-GeoData, the linked data version of OpenStreetMap. For each of the 34 POIs of this set, 5-6 photos depicting them were manually selected. In this way, a trusted set of 196 links which relates the POIs with their respective images was built. A higher number of photos of Milano POIs were collected from Wikimedia Commons and from Flickr. This second set of information consists of more than 37,000 candidate links that relate the POIs with the images that potentially depict them. This link-set is considered uncertain or untrusted. Those candidate links are expressed as RDF links using the foaf:depiction predicate and are further annotated with a confidence value that expresses the lack of certainty about their trustworthiness (e.g., the initial confidence of links to Wikimedia images is set to 60%, links to Flickr to 40%). The game UrbanMatch

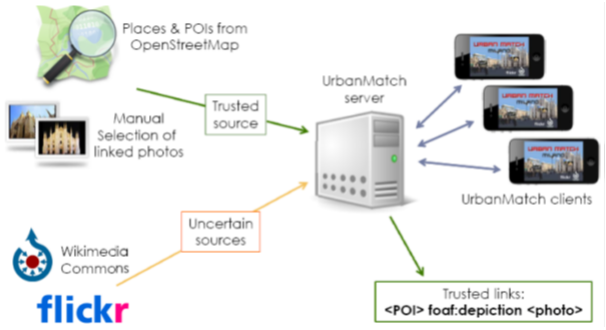

Fig. 3. Input data sources and output links in UrbanMatch [24].

is simply a photo coupling game. Eight photos of POIs are displayed; Four of these photos must come from a trusted source. Two must be taken from the set of candidate links, and thus are uncertain. Two are distractors in order to check the reliability of players. In this paper, authors succeeded to prove that mobile gaming applications can be successfully employed to consume, create and improve urban-related Linked Data, creating high-quality links between existing datasets,

(namely OpenStreetMap, LinkedGeoData and Flickr). As a conclusion we can see that different datasets can be created or improved via Human Computation approaches in the case of trade services, tourism, traffic optimization, environmental sustainability.

One last use case we depict in this survey is related to the rapid development in *biomedical research* which produces a continuous stream of new knowledge. On one hand, several open-access biomedical image portals are available on the internet (NBIA, NIH images, NCI visuals online, YALE image Finder). On the other hand, these systems have not published their contents in a semantically accessible manner. In [25], the authors present SEBI a semantic enrichment of biomedical images meta-data from YIF and enables search over that meta-data. For this purpose, this platform integrates a variety of best practice knowledge infrastructure components and services to generate sequence image annotations. The generation of these annotations is based on multi-modules each of which is responsible for generating a certain type of a sequence image annotation. When automatic annotation fails due to noisy input, annotation is made possible through the introduction of a crowd annotation technique. SEBI uses Semantic

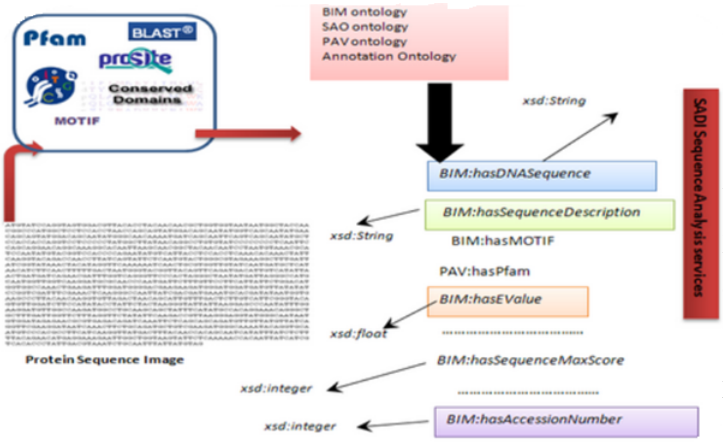

Fig. 4. Semantic Enrichment of a Sequence Image [25].

Automatic Discovery and Integration (SADI) that allows the integration and interoperability of resources by using Semantic Web standards. The annotations received from SADI web services are transformed into linked open data along with their respective images and are kept in triplestore called iCryus that can be queried through a SPARQL endpoint or navigated via a RDF browser. Fig 4 illustrates the semantic enrichment of an image. Finally, SEBI provides a portal through which a user can retrieve images based on semantic annotations.

IV. DISCUSSION

The projects that were detailed in III, and summarized in table I, are analyzed and compared according to the following characteristics:

1) **Domain**: The problems of data integration that have been exposed in the current study arise each in a very distinctive domain e.g. geoscience research, sensor Data, urban environments and biomedical image.
2) **Dataset**: Each of these projects treats very diverse data sources, ranging from structured (Public transport timetables, professional society membership, Job density,) semi-structured (Yale Image Finder, OpenStreetMap,) and unstructured (Flicker, conference abstracts, social media,).
3) **Semantic techniques**: Among these projects, the LOD is a common technique that is used to integrate and access data at various levels of complexities. Additional techniques are also used:
   - Semantic Provenance, PAV ontology and Vocabulary of Interlinked Datasets for metadata and annotation.
   - Semantic Automatic Discovery and Integration (SADI) for data integration.
   - Linked Stream Middleware, Semantic Sensor Network for sensor data integration.
   - Ontology Design Patterns for ontology alignment.
4) **Crowdsourcing techniques**: The role of crowdsourcing varies according to the case study issue. It was used for data validation in Ocean sciences, for data integration in Urban-Match and Biomedical image, and for data annotation in Geoscience research. The notable difference is with Urban-Match project [24], which identifies the mechanics used to enhance the crowdsourcing quality.

We can deduce that each one of the projects adopts different techniques in order to accomplish its objective of data integration. The main common thing is the use of Linked Data technology as an information infrastructure for integration of data from disparate sources. Furthermore, the output of all the studied integration processes is a linked open data repository.

It's clear that Linked Open data can solve some of the problems of Big Data integration and reduce heterogeneity aspect of Big Data. However, the case studies demonstrate that the integration of linked open data is challenged by the several quality issues and problems that the linked data paradigm is facing. Many of these quality issues require crowdsourcing techniques to be solved. As a result, all these case studies combine the use of LOD as a semantic web technology with different crowdsourcing techniques to solve the big data variety. We can see that:

- Geosciences [19] uses Crowdsourcing Entity Linking to link two entities with a new relation.
- Ocean link [22] uses Crowdsourcing Ontology Alignment to specify the correspondences between ontology entities.
- Linked Sensor Framework [23] uses crowdsourced data enrichment to clean and enrich data for accuracy.
- UrbanMatch [24] uses Game With a Purpose to link points of interests with the most representative photos.
- Biomedical Image [25] uses Crowdsourcing Image Annotation to allow users to annotate, delete, or update annotations.

Even though different techniques were used in these exposed

TABLE I
SUMMARY OF SELECTED EXISTING WORKS IN BIG DATA INTEGRATION USE CASES WHERE SEMANTICS MEET CROWDSOURCING.

Reference number	Domain	Dataset	Semantic techniques	Crowd Techniques
[19]	Geoscience research	Conference attendees, co-authorship, professional society membership, meetings attended.	Linked Open Data. Semantic provenance.	Crowd annotation. Portal to link data.
[22]	Ocean sciences	Ocean science data repositories, library holdings, conference abstracts, funded research awards.	Ontology Design Patterns. Vocabulary of Interlinked Datasets. Linked Open Data.	Portal to validate links
[23]	Sensor Data	Social data, Sensor data, Open data, Legacy data.	Linked Stream Middleware. Semantic Sensor Network.	Spatial crowdsourcing
[24]	Urban environments	OpenStreetMap. Wikimedia Commons. Flicker.	RDF links annotated with a confidence value. Linked Open Data.	Game with a Purpose to validate links.
[25]	Biomedical image	Yale Image Finder (YIF). PubMed. PDB. DrugBank.	Semantic Automatic Discovery and Integration (SADI). Linked Open Data. PAV ontology.	Crowd annotation

case studies, we succeeded in identifying a common workflow that could be used in any scenario, as we will see in more detail in the next section.

V. PROPOSED WORKFLOW

In this section, We detail the general workflow (see fig. 5) that could solve the problem of integration of heterogeneous big data by using semantic technology and crowdsourcing. The implementation of this workflow encompasses the following main steps: data preparation, resource selection, data modeling, data instantiation and finally data linking. To address the quality issues of this automated process, the workflow is extended with a parallel crowdsourcing process. In short, semantic and crowdsourcing analysis are applied to the heterogeneous data at the entry of the workflow and the resulting extracted entities are stored in a knowledge base, inter-linked with linked open data resources. In more detail, the steps are:

1) **Data preparation**: Raw data may be dirty, inconsistent, or incomplete. Thus, data preparation is the process of preparing (or pre-processing) data sources into refined information, which can be used effectively in data integration process. The data in this step need to be explored, organized, cleaned and augmented.
2) **Resource selection**: Choosing a target Linked Dataset to link with, requires extensive research work and a good analysis of the data source as well as the selected target.
3) **Semantic data modeling**: The main common goal in developing an ontology is to share the domain knowledge of the structure of information. A lot of work has been put into the investigation of ontology alignment. In general, this topic can best be treated under three headings: ontology mapping, ontology merging, and ontology matching. All of them are significant methodologies in managing semantic heterogeneity.
4) **Semantic data instances assessment**: Semantic data instances may enclose logical inconsistencies, and syntax errors. In other cases, these data instances can be incomplete and need to be enriched with additional information.
5) **Semantic data linking**: Linking data with LOD brings down the fences between various sources, but these links need to be created and validated.

As we can see these tasks may not be fully automated. Therefore, the support of crowdsourcing at each level can lead to an undeniable improvement of the quality of data and consequently of the quality of data integration outcome. More details at the conceptual as well as technical level will be elaborated in our future studies and implementations, primarily in investigating recent researches which merge crowdsourcing and NLP techniques.

VI. CONCLUSION

In this survey, we explored the steps needed for merging semantic and crowdsourcing technologies in order to enhance big data integration process and tackle analysis needs and challenges. Moreover, we clarified how the choice of a suitable approach for data integration depends on the types of Datasets used. In addition, we presented a conceptualization of the high-level data integration workflow summarized in a flow diagram (fig 5), where, for each step, we showed how extending the automated tasks with crowdsourcing would improve the quality of integrated data.

We found that by jointly using semantic and crowdsourcing technologies we have the possibility to address the big data variety issue. Indeed, these technologies provide necessary mechanisms to enable integration of different Big Data sets while ensuring quality, correctness and consistency. In our opinion, future work in this area must focus particularly on providing a generic framework based on the synergy between the Semantic Technologies and crowdsourcing to solve problems of big data effectively and efficiently. In fact, this study will be the used in support to the implementation of such a framework, in the context of relations extraction and validation from ”a semantic data lake”, in order to improve the quality of data integration and gain insights from various data sources.

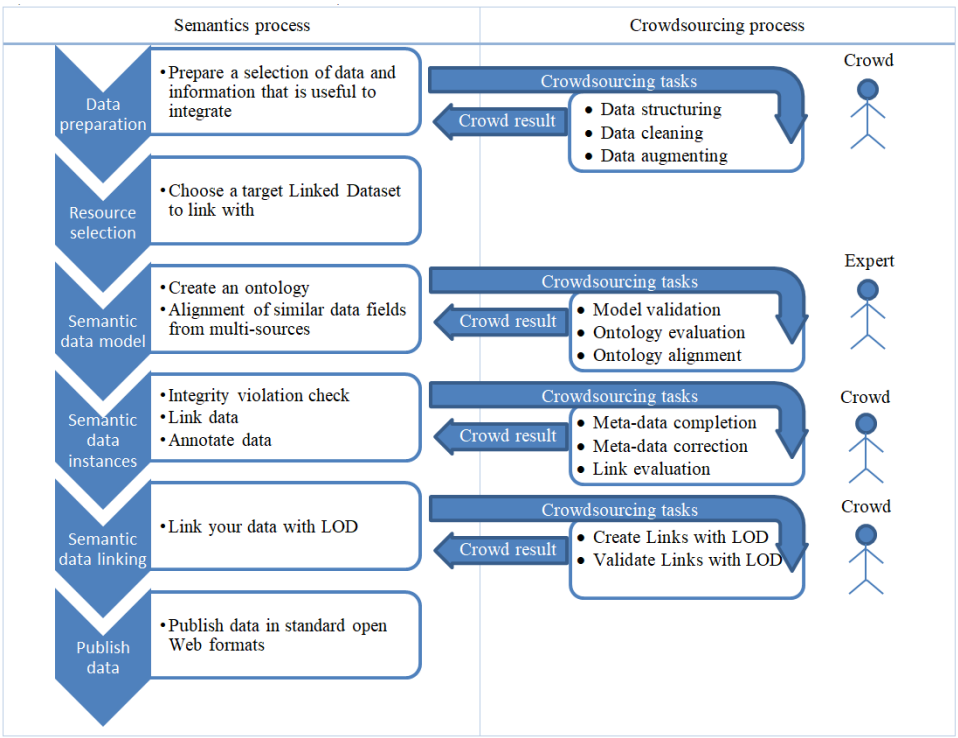

Fig. 5. A high-level overview of the proposed data integration workflow.

REFERENCES

[1] V. Prajapati, *Big data analytics with R and Hadoop*. Packt Publishing Ltd, 2013.

[2] E. Dumbill, J. Howard, M. Zwemer, M. Loukides, M. Slocum, A. Croll, J. Steele, C. Hill *et al.*, "Big data now-2012 edition," 2013.

[3] A. Labrinidis and H. V. Jagadish, "Challenges and opportunities with big data," *Proceedings of the VLDB Endowment*, vol. 5, no. 12, pp. 2032–2033, 2012.

[4] Y. Demchenko, P. Grosso, C. De Laat, and P. Membrey, "Addressing big data issues in scientific data infrastructure," in *Collaboration Technologies and Systems (CTS), 2013 International Conference on*. IEEE, 2013, pp. 48–55.

[5] X. L. Dong and D. Srivastava, "Big data integration," *Synthesis Lectures on Data Management*, vol. 7, no. 1, pp. 1–198, 2015.

[6] B. Arputhamary and L. Arockiam, "Data integration in big data environment," *Bonfring International Journal of Data Mining*, vol. 5, no. 1, p. 1, 2015.

[7] I. G. Terrizzano, P. M. Schwarz, M. Roth, and J. E. Colino, "Data wrangling: The challenging yourney from the wild to the lake." in *CIDR*, 2015.

[8] N. Heudecker and A. White, "The data lake fallacy: All water and little substance," *Gartner Report G*, vol. 264950, 2014.

[9] N. Shadbolt, T. Berners-Lee, and W. Hall, "The semantic web revisited," *IEEE intelligent systems*, vol. 21, no. 3, pp. 96–101, 2006.

[10] L. Mendoza, "A rule-based approach to address semantic accuracy problems on linked data," *ISWC-DC 2014 Doctoral Consortium at ISWC 2014*, p. 64.

[11] N. W. Paton and A. A. Fernandes, "Crowdsourcing feedback for pay-as-you-go data integration," *DBCrowd 2013*, p. 32, 2013.

[12] S. R. Jeffery, L. Sun, M. DeLand, N. Pendar, R. Barber, and A. Galdi, "Arnold: Declarative crowd-machine data integration." in *CIDR*, 2013.

[13] A. Benedek, G. Molnár, and Z. Szűts, "Practices of crowdsourcing in relation to big data analysis and education methods," in *Intelligent Systems and Informatics (SISY), 2015 IEEE 13th International Symposium on*. IEEE, 2015, pp. 167–172.

[14] D. C. Brabham, *Using crowdsourcing in government*. IBM Center for The Business of Government, 2013.

[15] C. Sarasua, E. Simperl, N. Noy, A. Bernstein, and J. M. Leimeister, "Crowdsourcing and the semantic web: A research manifesto," *Human Computation (HCOMP)*, vol. 2, no. 1, pp. 3–17, 2015.

[16] M. Acosta, E. Simperl, F. Flöck, M.-E. Vidal, and R. Studer, "Rdf-hunter: Automatically crowdsourcing the execution of queries against rdf data sets," *arXiv preprint arXiv:1503.02911*, 2015.

[17] A. Bernstein, "The global brain semantic web interleaving human-machine knowledge and computation," 2012.

[18] D. DiFranzo and J. Hendler, "The semantic web and the next generation of human computation," in *Handbook of Human Computation*. Springer, 2013, pp. 523–530.

[19] T. Narock and P. Hitzler, "Crowdsourcing semantics for big data in geoscience applications," in *AAAI 2013 Fall Symposium Series, Semantics for Big Data, November*, 2013, pp. 15–17.

[20] R. C. Roemer and R. Borchardt, "From bibliometrics to altmetrics: A changing scholarly landscape," *College & Research Libraries News*, vol. 73, no. 10, pp. 596–600, 2012.

[21] J. L. Fleiss, "Measuring nominal scale agreement among many raters." *Psychological bulletin*, vol. 76, no. 5, p. 378, 1971.

[22] T. Narock, R. Arko, S. Carbotte, A. Krisnadhi, P. Hitzler, M. Cheatham, A. Shepherd, C. Chandler, L. Raymond, P. Wiebe *et al.*, "The oceanlink project," in *Big Data (Big Data), 2014 IEEE International Conference on*. IEEE, 2014, pp. 14–21.

[23] D. N. Crowley, M. Dabrowski, and J. G. Breslin, "Decision support using linked, social, and sensor data," 2013.

[24] I. Celino, S. Contessa, M. Corubolo, D. Dell'Aglio, E. Della Valle, S. Fumeo, T. Krüger, and T. Krüger, "Urbanmatch-linking and improving smart cities data." in *LDOW*, 2012.

[25] S. A. C. Bukhari, M. Krauthammer, and C. J. Baker, "Sebi: An architecture for biomedical image discovery, interoperability and reusability based on semantic enrichment." in *SWAT4LS*, 2014.