Methods for Human-Machine Link Quality Management on the Web of Data

by
Cristina Sarasua Garmendia

Approved Dissertation thesis for the partial fulfillment of the requirements for a
Doctor of Natural Sciences (Dr. rer. nat.)
Fachbereich 4: Informatik
Universität Koblenz-Landau

Chair of PhD Board:	Prof. Dr. Ralf Lämmel
Chair of PhD Commission:	Prof. Dr. Maria Wimmer
Examiner and Supervisor:	Prof. Dr. Steffen Staab
Examiner and Supervisor:	Prof. Dr. Matthias Thimm
Further Examiners:	Prof. Abraham Bernstein, Ph. D.

Date of the doctoral viva: November 26, 2021

Por ti, aita.
Dedicated to my father and my mother.

Abstract

Semantic Web technologies have been recognized to be key for the integration of distributed and heterogeneous data sources on the Web, as they provide means to define typed links between resources in a dynamic manner and following the principles of dataspaces. The widespread adoption of these technologies in the last years led to a large volume and variety of data sets published as machine-readable RDF data, that once linked constitute the so-called *Web of Data*. Given the large scale of the data, these links are typically generated by computational methods that given a set of RDF data sets, analyze their content and identify the entities and schema elements that should be connected via the links. Analogously to any other kind of data, in order to be truly useful and ready to be consumed, links need to comply with the criteria of high quality data (e.g., syntactically and semantically accurate, consistent, up-to-date). Despite the progress in the field of machine learning, human intelligence is still essential in the quest for high quality links: humans can train algorithms by labeling reference examples, validate the output of algorithms to verify their performance on a data set basis, as well as augment the resulting set of links. Humans —especially expert humans, however, have limited availability. Hence, extending data quality management processes from data owners/publishers to a broader audience can significantly improve the data quality management life cycle.

Recent advances in human computation and peer-production technologies opened new avenues for human-machine data management techniques, allowing to involve non-experts in certain tasks and providing methods for cooperative approaches. The research work presented in this thesis takes advantage of such technologies and investigates human-machine methods that aim at facilitating link quality management in the Semantic Web. Firstly, and focusing on the dimension of *link accuracy*, a method for crowdsourcing ontology alignment is presented. This method, also applicable to entities, is implemented as a complement to automatic ontology alignment algorithms. Secondly, novel measures for the dimension of *information gain* facilitated by the links are introduced. These entropy-centric measures provide data managers with information about the extent the entities in the linked data set gain information in terms of entity description, connectivity and schema heterogeneity. Thirdly, taking Wikidata —the most successful case of a linked data set curated, linked and maintained by a community of humans and bots— as a case study, we apply descriptive and predictive data mining techniques to study participation inequality and user attrition. Our findings and method can help community managers make decisions on when/how to intervene with user retention plans. Lastly, an ontology to model the history of crowd contributions across marketplaces is presented. While the field of human-machine data management poses complex social and technical challenges, the work in this thesis aims to contribute to the development of this still emerging field.

Zusammenfassung

Semantic-Web-Technologien haben sich als Schlüssel für die Integration verteilter und heterogener Datenquellen im Web erwiesen, da sie die Möglichkeit bieten, typisierte Verknüpfungen zwischen Ressourcen auf dynamische Weise und nach den Prinzipien von sogenannten Dataspaces zu definieren. Die weit verbreitete Einführung dieser Technologien in den letzten Jahren führte zu einer großen Menge und Vielfalt von Datensätzen, die als maschinenlesbare RDF-Daten veröffentlicht wurden und nach ihrer Verknüpfung das sogenannte *Web of Data* bilden. Angesichts des großen Datenumfangs werden diese Verknüpfungen normalerweise durch Berechnungsmethoden generiert, den Inhalt von RDF-Datensätzen analysieren und die Entitäten und Schemaelemente identifizieren, die über die Verknüpfungen verbunden werden sollen. Analog zu jeder anderen Art von Daten müssen Links die Kriterien für Daten hoher Qualität erfüllen (z. B. syntaktisch und semantisch genau, konsistent, aktuell), um wirklich nützlich und leicht zu konsumieren zu sein. Trotz der Fortschritte auf dem Gebiet des maschinellen Lernens ist die menschliche Intelligenz für die Suche nach qualitativ hochwertigen Verbindungen nach wie vor von entscheidender Bedeutung: Menschen können Algorithmen trainieren, die Ausgabe von Algorithmen in Bezug auf die Leistung validieren, und auch die resultierenden Links erweitern. Allerdings sind Menschen – insbesondere erfahrene Menschen – nur begrenzt verfügbar. Daher kann die Ausweitung der Datenqualitätsmanagementprozesse von Dateneigentümern/-verlegern auf ein breiteres Publikum den Lebenszyklus des Datenqualitätsmanagements erheblich verbessern.

Die jüngsten Fortschritte bei Human Computation und bei Peer-Production-Technologien eröffneten neue Wege für Techniken zur Verwaltung von Mensch-Maschine-Daten, die es ermöglichten, Nicht-Experten in bestimmte Aufgaben einzubeziehen und Methoden für kooperative Ansätze bereitzustellen. Die in dieser Arbeit vorgestellten Forschungsarbeiten nutzen solche Technologien und untersuchen Mensch-Maschine-Methoden, die das Management der Verbindungsqualität im Semantic Web erleichtern sollen. Zunächst wird unter Berücksichtigung der Dimension *der Verbindungsgenauigkeit* eine Crowdsourcing Methode zur Ontology Alignment vorgestellt. Diese Methode, die auch auf Entitäten anwendbar ist, wird als Ergänzung zu automatischen Ontology Alignment implementiert. Zweitens werden neuartige Maßnahmen zur Dimension *des Informationsgewinns* eingeführt, die durch die Verknüpfungen erleichtert werden. Diese entropiezentrierten Maßnahmen liefern Datenmanagern Informationen darüber, inwieweit die Entitäten im verknüpften Datensatz Informationen in Bezug auf Entitätsbeschreibung, Konnektivität und Schemaheterogenität erhalten. Drittens wenden wir Wikidata - den erfolgreichsten Fall eines verknüpften Datensatzes, der von einer Gemeinschaft von Menschen und Bots kuratiert, verknüpft und verwaltet wird - als Fallstudie an und wenden deskriptive und prädiktive Data Mining-Techniken an, um die Ungleichheit der Teilnahme und den Nutzerschwung zu untersuchen. Unsere Ergebnisse und Methoden können Community-Managern helfen, Entscheidungen darüber zu treffen, wann/wie mit Maßnahmen zur Nutzerbindung eingegriffen werden soll. Zuletzt wird eine Ontologie zur Modellierung der Geschichte der Crowd-Beiträge auf verschiedenen Marktplätzen vorgestellt. Während der Bereich des Mensch-Maschine-Datenmanagements komplexe soziale und technische Herausforderungen mit sich bringt, zielen die Beiträge dieser Arbeit darauf ab, zur Entwicklung dieses noch aufstrebenden Bereichs beizutragen.

Acknowledgments

I am extremely grateful to many people for their support, collaboration and mentorship. I had the opportunity to work with scientists who taught me so much about scientific research.

First of all, I would like to thank Steffen Staab for his guidance, constructive feedback, and support. I would also like to thank all my colleagues at the University of Koblenz-Landau for their camaraderie and knowledge exchange, especially Matthias Thimm, who guided me through the PhD process, and Olaf Görlitz, who proofread chapters of the thesis and encouraged me towards the submission of this document.

Secondly, I would also like to thank all my other co-authors for the gratifying collaboration, in particular Natasha Noy and Elena Simperl, who not only triggered my interest in various scientific topics, but also became reference examples of women in science; Alessandro Checco, Gianluca Demartini, Djellel Difallah, and Michael Feldman, from whom I learned so much about data analytics and the joy of research; and Lydia Pintscher, who inspired me to be even more passionate about Wikidata.

Thirdly, I would like to thank Abraham Bernstein, who gave me endless support and from whom I gather valuable knowledge every day.

Moreover, I sincerely thank Claudia Müller-Birn for the encouragement and refreshing conversations.

Finally, I would like to deeply thank my family and friends. There are many, many other people who provided me with insights, experience and support that were key in this research and life path. Thanks to all of them as well.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

"The whole is greater than the sum of its parts." — **Aristotle**

Over the last decades, and especially with the emergence of the World Wide Web, there has been a substantial rise in the amount of available data. Open Data initiatives promoted by governments, research institutions and non-profit organizations such as Wikimedia Foundation[1] and Open Knowledge Foundation[2], have succeeded in raising awareness about the importance of publishing so-called *open data* to facilitate its usage and boost organization and government transparency. As a result, many of the available datasets are published to be consumed and reused by anyone for free[3]. The Open Data Monitor[4] identified, as of February 2020, 1.4+ TB of data licensed as open from 173 data catalogs of many different countries. The English Wikipedia and Wikidata have grown up to 5.9 million articles and 68 million items respectively, licensed as public domain under a Creative Commons license CC0[5].

This vast amount of data describing different domains and stored in data sources that operate independently, has materialized in the shape of both structured and unstructured data published in, for instance, HTML Web sites, PDF text documents, relational databases, CSV text files, XML documents, and JSON data files. Private companies store data about their products and corresponding reviews, as well as business transactions, user profiles, product usage statistics and service performance logs; government organizations publish statistical data about population growth, employment, environment and other topics relevant for society; archival institutions maintain digital bibliographic datasets, and research scientists produce lots of empirical data. There is, hence, a ***variety*** of data generated by different agents for various purposes that presents a high level of heterogeneity and is distributed over different physical locations.

There are many situations in which consuming data coming simultaneously from various distributed and heterogeneous datasets becomes useful. Let us think of a research data management scenario, where individual organizations with a common research end-goal collect and analyze data independently. Their datasets can be complementary (e.g., different information about the same individuals, information about different but analogous individuals) and consuming them at once, as if they belonged to a single database, can help answer more complex queries (e.g. queries that use more attributes for

[1] WMF `https://wikimediafoundation.org/wiki/Home`
[2] OKFN `https://okfn.org/`
[3] Open Definition `http://opendefinition.org/`
[4] Open Data Monitor `http://opendatamonitor.eu/`
[5] CC0 `https://creativecommons.org/publicdomain/`

the data selection), retrieve more or more accurate results and facilitate systematic data comparisons for quality assurance purposes. This process of "combining data residing at different sources, and providing the user with a unified view of these data" receives the name of *data integration*, as stated by Lenzerini (2002). Materializing this data integration entails aligning the individual datasets, explicitly specifying the way they relate to each other, and providing the necessary infrastructure to allow a seamless execution of multi-dataset queries. This process has been studied for decades (Wiederhold, 1992; Ziegler and Dittrich, 2004). In the technological context of this thesis (Semantic Web technologies), this alignment between the different datasets is specified through *links* that define the type of relationship between pairs of resources in each of the datasets.

Given the large scale of the data, there can be a large number of meaningful links between two datasets —hundreds of thousands. Hence, manually inspecting the datasets and specifying these links one by one often becomes an unfeasible option. For decades, the scientific community has studied algorithms that given a pair of datasets find the set of meaningful links between the resources of these two datasets, usually based on the relation of similarity (Christen, 2012). Despite the advance of AI technology and other computational approaches, there are often cases in which automatic algorithms do not reach perfect semantic accuracy —one of the most relevant data quality dimensions (Wang and Strong, 1996)— because the data is highly heterogeneous, or certain data values are missing. In other cases, algorithms are able to perform highly accurately but in order to do so they require a set of reference cases (i.e., training data) to learn from. For these reasons, there is a need to devise *solutions that combine human and machine intelligence to ensure the highest link quality possible*. Low quality links (e.g., semantically inaccurate links that specify that an entity of one dataset is the same as an entity of another dataset, when in the real world they do not refer to the same thing) can lead to misleading query results, as well as flawed data analysis, which could have disastrous consequences in many domains/applications for which data quality is critical for decision making.

The research presented in this dissertation is in the realm of human-machine link quality management in the Semantic Web. The overarching goal of the thesis is to provide methods that combining human and machine intelligence, help semantic data managers to obtain high quality links between their own datasets and other external datasets. In this work, different human-machine scenarios are studied: in some, humans aim to help machines, and in others machines aim to help humans. First, focusing on the link quality dimension of semantic accuracy, we investigate how crowdsourced human computation can help post-process the links produced by algorithms. Second, focusing on the link quality dimension of information gain facilitated by the links, we investigate novel measures that can be computed and presented to data managers to inform them better about the effect their links have. Third, in a scenario in which a community of humans and machines work collaboratively on the full semantic data (quality) management cycle, including the task of linking, we study how computational methods for analysis and behavior prediction can help identify which humans will be able to contribute to a larger extent.

1.1 Introduction to Data Integration in the Semantic Web

In order to provide the reader with the background information necessary to understand the challenges and contributions listed in this chapter, in this section I briefly introduce the technical evolution of the field of data integration, state-of-the-art methods to integrate data in the Semantic Web, as well as the notions of data and link quality in the Semantic Web.

Evolution of Techniques for Data Integration The task of data integration is particularly challenging because of (i) the *heterogeneity* (i.e., different formats, different data modelling conventions) and (ii) the *autonomy* (i. e. independent data management processes run by different agents) of the data in the individual sources (Doan et al., 2012). The problem of data integration has been broadly investigated in the field of databases (Doan et al., 2012), where initial contributions focused on materializing a global database with a single schema aligning the individual schemas of the heterogeneous databases (Thomas et al., 1990) (i. e. the individual datasets are extracted, transformed and loaded into a single data source). The main disadvantage of initial approaches is that data is duplicated, thus maintaining the global database up-to-date becomes a highly demanding task. The next generation of data integration solutions focused on a virtual integration implemented via a so-called mediated schema that defines the mappings between the individual datasets. Several methods were implemented to address this problem: "Global-As-View (GAV) modeled the global schema as a set of view definitions over the schemas of the data sources", whereas "Local-As-View (LAV) proposed to model the contents of a data source as a view over the global schema", and a combination of both was implemented in "Global-Local-As-View (GLAV) where a view of the data sources is defined as a view over the global schema." (Golshan et al., 2017; Doan et al., 2012). While each of these methods has a different query reformulation process (i. e. translation of a global query to individual data source queries), they are designed under a common assumption: that the mediated schema needs to be implemented a priori, considering all individual sources.

Over time, new data integration scenarios emerged (especially in the context of the Web), which needed a more flexible and dynamic integration process for larger amounts of data and more complex schemas. This led to the design of a new data integration paradigm introduced by Franklin et al. (2005); Halevy et al. (2006) named ***dataspaces***. In dataspaces, mappings are specified gradually, as needed over time —that is why this kind of data integration is often referred to as a "*pay-as-you-go* integration". This approach, in contrast to the more classic data integration methods, can handle many large datasets and facilitates the addition of new data sources to the integrated space. Furthermore, dataspaces are permissive with changes in the datasets and inconsistencies may co-exist (Golshan et al., 2017). In a parallel effort, the Semantic Web scientific community investigated methods and developed technology for annotating and linking Web resources in a machine-readable way, that extend the *Web of documents* into the so-called *Web of Data*. These technologies are indeed aligned with the principles of dataspaces (Heath and Bizer, 2011) and are very suitable for solving data integration problems in scenarios with broad heterogeneity and dynamic nature. With RDF (Resource Description Framework) (Antoniou et al., 2012), a metadata data model at the core of Semantic Web technologies, one can integrate distributed datasets by writing so-called RDF triples/statements that explicitly define the kind of relationship between pairs of concepts and pairs of entities that lay in the different datasets to be integrated. The vocabularies to describe such entities and relations can be reused and extended. The work presented in this dissertation focuses on this last option for data integration, based on Semantic Web technologies.

Interlinking in the Web of Data RDF statements are specified as triples of the shape "`subject predicate object`". These triples indicate some piece of information about an entity —the subject in the triple. Links are RDF triples where the subject and the object of the triple are the entities connected from the different datasets [6]. For instance, the triple "`gn:7287650 owl:sameAs`

[6]It is worthwhile mentioning that in the Linked Data paradigm, the separation between datasets is sometimes not well delineated: if there is no explicit dataset declaration (e. g. as a `void:Dataset` instance), nor a SPARQL endpoint setting

wd:Q72" is an identity link connecting the entity denoting Zurich in Geonames (dataset 1), and the entity for Zurich in Wikidata (dataset 2). One can also define links between concepts (i. e. at the schema level), which are also known as *ontology mappings*. An example of a schema mapping is the triple "o1:Company rdfs:equivalentClass o2:Organization". While the most widely-used predicates in RDF links refer to the relation of identity (Schmachtenberg et al., 2014), Semantic Web technologies enable the specification of any kind of relation, including subsumption, geolocation, and any other domain-specific relation. These relations are defined together with concepts and constraints in schemas (or vocabularies or ontologies); while link triples are stored in the datasets that primarily describe the entities in the subject position.

Linking to external entities is one of the core requirements for a published dataset to be considered as Linked Data (Berners-Lee, 2006). Hence, it is usually data publishers who generate links from their datasets to external datasets. In some cases, data publishers have concrete use cases and plan to consume their data in conjunction with external datasets (in an e. g. user application); other times, publishers expose their data for third-parties to use it, without a specific consumption goal in mind. In the latter case, publishers tend to aim at linking their datasets to so-called interlinking hubs (i. e. datasets that contain general-purpose data and a large number of entities popular across-domains, hence, relevant for many datasets in terms of identity linking) since that increases the visibility/discoverability of their data. Depending on the scale of the datasets, links can be created manually; however, this approach quickly becomes impractical with a medium-size dataset (i. e. starting, possibly, at several hundreds of entities to be connected). *Link discovery* frameworks (e. g. Silk, LIMES (Nentwig et al., 2017)) help data publishers in analyzing the descriptions of entities of the datasets to link and deciding programmatically what pairs of entities should be linked and what pairs should not. Analogously, *ontology alignment* frameworks (e. g. LogMap) analyse the definitions of concepts and properties to decide how (pairs) of ontologies should be aligned. The majority of these frameworks, operating at the schema- and/or data-level, are domain-agnostic, take as input two datasets (or two ontologies), and give as an output a set of identity or equivalence links (Nentwig et al., 2017; Shvaiko and Euzenat, 2013). Early frameworks were designed as batch linking tools, that required data publishers to define the necessary conditions (e. g. type of entities, value similarity for particular properties) for a pair of entities to be linked. Current state-of-the-art frameworks provide a supervised learning solution to the problem: the algorithm receives a reference set of links (i. e. a labeled link set) from the data publisher, and learns the rules that determine what is a link and what is not. The algorithm then applies such learned rules to the set of candidate links. In some of the frameworks, this process is implemented as an active-learning algorithm (Settles, 2012a), where the algorithm decides what candidate links need to be labeled, in order to minimize the number of labels to collect from the expert labeler. Chapter 2 provides further technical details on the foundations of link discovery technology.

Since the uptake of Semantic Web and Linked Data technologies, more than a decade ago, and the recent rise of knowledge graphs, many RDF datasets and vocabularies have been published; Semantic Web data management tools have evolved, and best practices have been refined. As a result of this activity, initially promoted by the scientific community and currently widely adopted by industry (Gaos et al., 2018), the public Web of Data contains (as of February 2020) a network of 1407 RDF graphs[7], including entity descriptions of various domains (e. g. geography, life sciences, publications, media),

a physical boundary, entities are part of the global Web dataspace. Entity descriptions can be obtained by dereferencing URIs, but URIs may be defined following different conventions. There is a certain understanding in the community, though, that datasets can be virtually distinguished by pay-level-domains (Schmachtenberg et al., 2014)

[7] Linked Open Data Cloud (LOD cloud) https://lod-cloud.net/,
Data Commons https://datacommons.org/

that are licensed under different conditions. As of February 2020, the RDF link repository LinkLion (Nentwig et al., 2014) captured 15.5 links published on the Web of Data, which used 12 different link predicates. By the same time, LODStats (Ermilov et al., 2016) computed a total of 33761 linked namespaces.

Semantic Web Data Quality and Link Quality Management For decades, the data and information quality literature has studied the meaning of *data quality*, as well as methods to assess it, monitor it and encourage its improvement. Traditionally, data quality was tightly connected to accuracy (i. e. the property that refers to data being semantically correct). Later on, it was acknowledged that data quality can be interpreted from different angles, and can thus have a multi-dimensional definition. Juran introduced the concept of "fitness for purpose" that emphasizes the importance of ensuring that the data fulfills the requirements of its intended goal (Juran, 1989). Wang and Strong (1996); Wang (1998) defined data quality as "data that are fit for use by data consumers". Wang and Strong surveyed data consumers in order to identify the attributes that they consider important when consuming the data. The result of their study was a framework for data quality dimensions grouped into 4 main categories: "intrinsic, contextual, representational and accessibility dimensions". Examples of intrinsic dimensions are *accuracy* (i. e. "the extent to which data are correct, reliable, and certified free of error"), and *objectivity* (i. e. "the extent to which data are unbiased (unprejudiced) and impartial"); contextual dimensions include *completeness* (i. e. "the extent to which data are of sufficient breadth, depth, and scope for the task at hand"), and *timeliness* (i. e. "the extent to which the age of the data is appropriate for the task at hand"); *representational consistency* (i. e. "the extent to which data are always presented in the same format and are compatible with previous data") and *ease of understanding* (i. e. "the extent to which data are clear without ambiguity and easily compreheneded") are representational dimensions; and *access security* (i. e. "the extent to which access to data can be restricted and hence kept secure") is an example for an accessibility dimension.

In the context of Semantic Web data, the widely-accepted data quality framework by Wang and Strong (1996) was adopted and extended in a literature review of quality assessment methods for Linked Data by Zaveri et al. (2016). One of the additional dimensions introduced in the Linked Data Quality article compared to the initial Data Quality article is *interlinking*. This dimension was primarily classified as an accessibility dimension, but also mentioned in the context of completeness. After all, links can be considered an excerpt of data, and can therefore be assessed in terms of the data quality dimensions, including, for example, accuracy and consistency. The link quality assessment procedures listed in the survey by Zaveri et al. (2016) mostly refer to checking if links to external entities exist at all, evaluating their accuracy either programmatically or manually, as well as determining the degree to which the interlinking of two datasets is fully materialized according to some notion of upper-bound completeness. The other additional dimension refers to data licensing.

The field of ontology engineering dedicated extensive efforts to defining ontology evaluation measures and methodologies (Vrandecic, 2010). Some of these works adopted quality dimensions from data quality frameworks (e.;g. consistency, completeness, conciseness), and extended them with ontology-specific dimensions, such as expandability, adaptability and mappability (i;e. whether an ontology "can be mapped to upper level or other ontologies") (Gómez-Pérez, 2004; Obrst et al., 2007).

All in all, there are many methods to define and assess quality, but there are fewer methods that focus on the actual improvement of data quality.

1.2 Motivation Scenario

In order to better showcase state-of-the-art Semantic Web technologies for data integration, as well as to motivate the challenges described below, let us imagine a research data management scenario: three different institutes investigate user behavioral patterns in peer-production systems. All of them conducted location-based studies and gathered research datasets that provide measurements for each of the users in the system they observed. The three datasets have in common the description of some users (who have contributed to several of the studied systems), and all of them report measurements for some frequently used behavior indicators. However, the datasets present the following heterogeneities:

- they contain different (meta)data (e. g. in one dataset, basic user profiles contain information about `age`, `location` and `Big-Five personality trait`, while in other datasets user profiles consist of only `location`).
- researchers have described entities with different granularity (e. g. in one dataset the volume of user actions is computed on a weekly basis, while another dataset records daily user activity) and used different conventions (e. g. one dataset uses ISO country codes to refer to countries, while another dataset writes the complete name of countries).
- the datasets use multiple natural languages(e. g. some description texts are written in German, while others in English).
- the datasets have been published in various formats (e. g. CSV files, SQL databases and HTML Web pages), since each organization has its own technical infrastructure.
- besides factual data, two of the datasets also contain multimedia content such as maps and plots.

After noticing the relatedness of the datasets, Sarah —a researcher in one of the institutes— decided to integrate the datasets and conduct a comparative study using the data holistically. Given the characteristics of the problem at hand, integrating the data following the principles of dataspaces seems to be the most suitable technical option. In order to facilitate the further reuse of such integration and increase the FAIR-ness (Wilkinson et al., 2016) of the data, Sarah chose to implement mappings between entities in the datasets in a declarative manner, using Semantic Web technologies. Table 1.1 shows a set of resulting links both at a schema- ($l1$-$l5$) and data-level ($l6$-$l10$). Entities are uniquely identified by URIs (here in a shortened version with namespaces). As any information expressed in RDF, links are modelled as triples (i. e. an entity/concept from the source dataset/ontology in the `subject` position, the relation in the `predicate` position and an entity/concept from the target dataset/ontology in the `object` position). Sarah's dataset represented with `d1` and its ontology as `o1`. The mappings explicitly define the relation between concrete items of the datasets. For instance, the concept `Country` in `o1` is equivalent to the concept `Staat` in `o2`, and equivalent to the concept `Land` in `o3`; Sarah's ontology has more specific roles defined, hence, the administrator role is mapped to the concept `Editor` in `o2` with the relation of subsumption; edit actions in `o1` are part of `o3` routines, which aggregate actions that take place repeatedly; and the user indicator for `survival` in `o1` defined as a property, is equivalent to the indicator `lifespan` in `o3`. Regarding the data-level, the entity attributed to Germany in `d1` and the entity identified by the ISO country code `DE` refer to the same real-world entity; users who

registered with the same username across systems can also be mapped with `sameAs` relationships; the observed cluster of users in Sarah's dataset is linked to Wikidata's item for the city of Zurich, because Wikidata (the collaborative and multilingual free knowledgebase (Vrandecic and Krötzsch, 2014)) has useful statistical data about the population of the city over time, that could be used to contextualise some findings; an entity formally describing a debate that took place online is typed as a `sioc:Post`, a concept defined in the external and standard de facto ontology for online communities; finally, the plot in `d1` illustrating the editing evolution of users in System 1 is connected to the analogous chart about System 2, using a more general link predicate (`rdfs:seeAlso`), which vaguely suggests to look up the entity in the object position when encountering the entity in the subject position.

example	source	predicate	target
l1	o1:Country	owl:equivalentClass	o2:Staat
l2	o1:Country	owl:equivalentClass	o3:Land
l3	o1:Admin	rdfs:subClassOf	o2:Editor
l4	o1:EditAction	dct:partOf	o3:Routine
l5	o1:survival	rdfs:equivalentProperty	o3:lifespan
l6	d1:Germany	owl:sameAs	d2:DE
l7	d1:U12908	owl:sameAs	d3:User98
l8	d1:CommClusterCK	gn:locatedIn	wd:Q72
l9	d1:Debate18072017	rdf:type	sioc:Post
l10	d1:Plot1A	rdfs:seeAlso	d2:ChartEvU1

Table 1.1: Examples of schema mappings (l1-l5) and data links (l6-l10) in a research data management scenario.

Aligning concepts and entities this way, it is possible to jointly consume the data, being able to retrieve user data collected in partially in the different datasets, and using queries with filtering conditions that apply to all datasets (e. g. we can ask to filter per country and equivalent classes, independently of whether they were defined in a different language or with different properties). These triples become part of Sarah's dataset, and hence data consumers can query her and the other two datasets using the SPARQL query language in a federated fashion[8].

To obtain some of the links in Table 1.1, Sarah used a so-called *ontology mapping* tool to align the ontologies and a *link discovery* tool to create the links between data entities. Such tools operate similarly and assume that the data publisher has a pre-design of the expected outcome in mind, at least in terms of the desired type of entities to be connected and the link predicates to be used in the links. The vast majority of data-agnostic tools for this purpose focus on identity and subsumption links, but are not able to discover domain-oriented (e. g. `locatedIn` or less subtle links (e. g. `seeAlso`); consequently, for such cases Sarah had to create links either manually or using additional tools that apply specific domain rules to the link classification problem. In terms of accuracy, link discovery (and ontology mapping) tools exhibit high performance, yet not perfect (cf. Ontology Alignment Evaluation Initiative for ontology and entity matchers `http://oaei.ontologymatching.org/2019/`). Hence, given that data quality is important for Sarah, she inspected the links generated by the tools —to control for *precision*— and examined the datasets to identify missing links —to audit *recall*. In fact, the mapping resulting from an (semi-)automatic algorithm may contain false negatives (i. e. candidate links that are classified as non-link, when the ground truth indicates that it should be a link) and false

[8]SPARQL Federated Query `https://www.w3.org/TR/sparql11-federated-query/`

positives (i. e. candidate links that are classified as link, when the ground truth indicates that it should not be a link). For example, an algorithm can determine that two entities labelled after "TBL" and "Tim Berners-Lee" respectively do not refer to the same person (a false negative) if the dataset does not have additional information that can signal that TBL are the initials for Tim Berners-Lee and the algorithm decides on the classification based on uniquely the label. Analogously, the algorithm can classify two entities named after "Koblenz" as being the same entity, if it is not able to consider the fact that one is located in Germany and the other in Switzerland (a false positive).

1.3 Research Challenges

There is scarce empirical evidence of the quality of the data published in the Semantic Web, and more specifically, in the Linking Open Data Cloud; usually the focus of state-of-the-art investigations lays on the evaluation of methods over a sample of this data. A few studies partially analyzed the data in terms of its *conformance of the Linked Data principles* (Hogan et al., 2012), as well as its *volume* (Schmachtenberg et al., 2014), and *dynamics* (Käfer et al., 2013; Dividino et al., 2014). These works revealed, among other facts, that **(i)** the guidelines pertaining to documenting the creation and consumption of datasets were ignored to a larger extent than the guidelines that referred to technical recommendations on URI specification and maintenance; **(ii)** from 2011 until 2014 “the number of datasets approximately doubled and there was an increased agreement on common vocabularies for describing certain types of entities”; and **(iii)** approximately 62% of the documents did not change in content, while their availability was not close to ideal, since “documents were unavailable 20% of the time, from which 5% had gone permanently offline”.

Specifically regarding links, Schmachtenberg et al. (2014) found that ”there is still a relatively small number of datasets that set RDF links pointing at many other datasets, while many datasets only link to a few other datasets", “there are several category-specific linking hubs besides general-domain linking hubs such as DBpedia", and `owl:sameAs`, `rdfs:seeAlso` and `foaf:knows` are the most frequent link predicates — the latter in the context of social networks. Käfer et al. (2013) observed changes of 86+ thousand linked data documents for 29 weeks and observed that “the rate of fresh links being added to the documents is low”.

Recently, Färber et al. (2018) provided an extensive quality assessment on current linked data. The authors computed quality statistics about five major knowledge graphs —DBpedia, Freebase, OpenCyc, Wikidata, and YAGO— using both state-of-the-art and self-defined dimensions/measures, together with gold standard data they curated. The results of their study show that in some dimensions all knowledge graphs had a similar positive score (e. g. in syntactic validity, and semantic validity of triples), while in others, some of the knowledge graphs scored better than others (e. g. Wikidata had, at the time of the analysis, the highest schema completeness; whereas OpenCyc and Wikidata were better than the other knowledge graphs in terms of trustworthiness). There are measures that are unique to some of the knowledge graphs (e. g. only Wikidata has the feature of statements ranking). This study shows that data consumers may want to decide which knowledge graph to consume based on the dimension they value or need the most in their concrete scenario. Actually, Färber et al. (2018) provide in their article a framework for ranking the knowledge graph’s content, given a set of user-defined weights that indicate the relevance of each linked data quality dimensions. One of the limitations of this study, though, is that dimensions such as semantic validity and schema completeness were not analyzed comprehensively throughout the full knowledge graphs: as explained by the authors, they evaluated such dimensions over a small data sample (e. g. semantic validity of triples was calculated over 100 people

instances, and 400 facts per knowledge graph), because such dimension assessments require expert-curated gold standards and they are difficult to create. A complementary study accomplished by Debattista et al. (2018), applied 27 quality measures from the related work and two self-defined measures to 130 LOD datasets. Some of the measures the authors used have common ground with the measures used by Färber et al. (2018); for example, they both look at incorrect typing, and violated disjointness constraints. Nevertheless, Debattista et al. (2018) did not include semantic accuracy and completeness assessments. An interesting contribution of this work, is that the authors applied the PCA technique over the set of quality measurements they computed, in order to investigate which measures provide redundant information in terms of quality. The authors found that in their experiments, 3 measures "(links to external data providers, usage of incorrect domain or range datatypes and dereferenceability) provide no new information" compared to the information provided by the other quality measures.

A well-known and persistent quality problem in the Web of Data, is the existence of semantically incorrect `owl:sameAs` links. Halpin et al. (2010) identified several misuses of the identity predicate, whose intended use is to specify that two entities are exactly the same real-world thing. Halpin et al. identified multiple misuses of this predicate in RDF data, based on individual Linked Data observations: (i) two entities might be the same real-world thing but "all the properties ascribed to one URI are not necessarily accepted by the other URI"; (ii) two URIs refer to the same thing but they should be reused in different contexts (e. g. it might not desirable to use the work-related description of a person in an informal context and vice versa); (iii) the relation refers to representation instead of identity (e. g. one of the URIs is the Web site of a person but it is not an entity referring to the person per se); (iv) the relation refers to a close similarity, but not really identity; (v) the two entities might be closely related by domain, but they are not identical. With this in mind, Halpin et al. proposed the use of alternative predicates defined in various vocabularies, such as `skos:closeMatch`, `skos:narrowMatch`, `skos:broadMatch` and `foaf:primaryTopicOf`, as well as the use of named graphs to define contextual boundaries, to specify the meanings in the aforementioned cases more precisely. Furthermore, the authors introduced the similarity ontology containing eight similarity properties that "are not necessarily transitive and symmetric", and aligned to existing properties specified in RDFS, OWL and SKOS. As indicated by Raad et al. (2019), the amount of erroneous identity links was estimated to be "20% [of all published `owl:sameAs` links] in 2010 according to Halpin et al. (2010) and 2.8% in 2012 according to Hogan et al. (2012)". Considering that identity links are to a large extent responsible for building the backbone of the Semantic Web dataspace, the situation is alarming.

In terms of schema mappings, there are fewer studies examining the quality of existing aligned ontologies. Asprino et al. (2019) studied a cached version of the LOD data, including over six hundred datasets (dubbed LOD-A-Lot [9]). The authors analyzed characteristics of the alignment and the hierarchy of the ontologies in the individual datasets, and found that concepts in these ontologies, and even more so properties, are "sparsely linked with equivalence relations".

In an attempt to improve the quality of current links, the community considerably pushed forward methods that aim at identifying erroneous identity links programmatically. As summarized by Raad et al. (2019), some of these methods analyze the meaning of existing links to detect inconsistencies in the data, caused by the interaction between `owl:sameAs` and `owl:differentFrom` statements, or any other ontological axioms (e. g. symmetry constraints etc.). Another set of works solved this task with the application of network analysis techniques, using measures like centrality or community partitioning (Guéret et al., 2012; Raad et al., 2018). Others used outlier detection methods (e. g. based on clustering and cosine similarity) (Paulheim, 2014). These methods provide a very useful approach for

[9] LOD-A-Lot `http://lod-a-lot.lod.labs.vu.nl/`

an initial unsupervised quality screening, but given that their accuracy leaves room for improvement, we still heavily rely on human action.

These statistics show that there is a clear need for procedures and frameworks that facilitate more effective link quality assurance. As acknowledged by various (linked) data and information quality methodologies (Rula and Zaveri, 2014; Batini et al., 2009; Lee et al., 2002), quality assurance goes beyond quality assessment: it should also involve assessment-improvement cycles in which fine-grained and multi-dimension quality issues are identified, corrected and enhanced. Hence, to improve the current link quality situation, it is important to investigate methods that enable continuous quality assurance supported by humans.

The research work presented in this thesis addresses four major challenges that arise specifically in the context of human-machine link quality management. Firstly, we focus on facilitating large-scale human computation for ensuring semantic accuracy in ontology alignment, one of the major tasks in semantic dataset integration. The second challenge we address refers to assessing and improving link quality beyond well-studied dimensions, such as accuracy and (relative) completeness, in order to further align the notion linked data quality to one of the core Linked Data principles. Thirdly, and moving one level of abstraction higher to the scenario of general-purpose human-machine knowledge engineering, we concentrate on human users' contributions, retention management and recognition for work, in the context of real-world crowd-powered systems.

1.3.1 Facilitating Large-Scale Human Computation for Link Accuracy Management at a Schema-Level

Linking datasets in the Web of Data may be done exclusively at a data-level (i.,e., without integrating the schemas used to annotate the entities in the datasets). In fact, the LOD movement focused to a large extent on connecting entities. However, integrating the schemas provides additional benefits in terms querying the data and administering its quality. For example, an instance may be included in some query results precisely because of a type inherited from its linkage to another instance.

Past benchmarking campaigns [10] showed that purely automatic solutions to ontology alignment still outperform hybrid human-machine solutions. While there might exist the perception that ontologists should be able to manually process ontology mappings and assess their semantic accuracy, because their volume is presumably smaller than data links, that is very often not the case. Projects in domains such as biomedicine led to large-scale ontologies with hundreds of thousands of classes [11]. Moreover, sometimes gold standard datasets need to be created (to serve as ground truth in evaluations). Hence, it is necessary to provide means that can systematically outsource to humans the computational process of classifying a large set of candidate pairs of concepts/properties as equivalent or subsumed.

Vrandecic (2010) proposed how to evaluate ontologies collaboratively in a peer production fashion (Benkler et al., 2015) within Semantic MediaWiki, a MediaWiki extension that is capable to annotate knowledge in a wiki in terms of ontology-based metadata (Krötzsch and Vrandecic, 2011). The system automatically checks constraints on class disjointness, concept and property cardinality, and the community constituted around such a wiki may fix and enhance the knowledge specification based on the tested constraints. In such a system, users can also collaborate in the creation and review of ontology mappings. Such a community of users can have an expert profile (e. g. neuroscientists and psychologists trying to establish a common terminology) but as mentioned, there are typically more mappings to process than dedicated humans to do so. For projects in which building a community

[10] Ontology Alignment Evaluation Initiative `http://oaei.ontologymatching.org/`
[11] BioPortal ontologies `https://bioportal.bioontology.org/ontologies`

might be impractical, it is necessary to have a solution that offers a certain guarantee for human input (i. e. providing incentives that extrinsically motivate people to work on the task).

The state of the art has examples that demonstrate the utility of using microtask crowdsourcing for various tasks related to linking. For instance, Demartini et al. (2012) implemented a crowd-powered approach to entity linking method (i. e. identifying that a word in a text refers to a concrete entity defined in a knowledge graph). Their framework has a probabilistic component to decide which links from a set of links generated by an automatic NLP algorithm need further human inspection. Their evaluation showed that using microtask crowdsourcing considerably improves the output of the algorithm. Franklin et al. (2011) presented a database query processing solution that asks crowd workers to provide input (e. g. finding missing data and comparing data) to be able to solve queries that require user interaction because otherwise the query engine would return either empty or wrong results.

These works lay the foundation on the use of microtask crowdsourcing to solve data management problems. Notwithstanding, they do not prove the effectiveness of using this technique to solve the problem of quality assurance in ontology alignment, which is a more abstract task, as it refers to conceptual modeling instead of concrete entity data. Ontologies are frequently poorly documented, and crowdsourcing reaches out to people who may have an understanding of a domain/general human knowledge very distant from the ontology engineer(s) who initially designed the ontologies. Moreover, in ontology alignment the number of similar mapping cases can be small, which makes it harder for humans to get trained in a particular task. All in all, there is a need for a feasibility study that examines whether crowd workers are suitable for such a high-level task, and explores possible alternative solutions that allow for a combination of human and machine computation in ontology alignment.

1.3.2 Extending the Notion of Link Quality Management based on Information Gain

From all quality dimensions, *accuracy* is the dimension that received the most attention in the literature of link quality. Besides the related work on automatic methods for the identification of potential errors in identity links, more and more studies led to the implementation of human computation tasks. Kontokostas et al. (2013) built their own tool and formulated the solution as a community contest, while Acosta et al. (2013) defined microtasks to assess link triples via paid crowd work in Amazon Mechanical Turk. Ensuring that link statements are valid/true is, understandably, a critical matter for data to be consumed. However, if we aim for a more-interconnected Web of Data, to enable (i) richer federated query results and (ii) more extensive link path traversal (in breadth and depth), we need to measure and assess quality in different ways, as well as provide tools that encourage the design and creation of new and different links. After all, Tim Berners-Lee explicitly indicated in the Linked Data design principles that we should "include links to other URIs, so that they can discover more things".

Given a dataset and an existing set of outgoing links to entities of external datasets, further links may be created in different ways: we can (a) create links (with the same or a distinct predicate) to new datasets; (b) add the same type of links, to new entities of the same target datasets; and (c) add new types of links to the same entities of the same target datasets. The decision in any of these cases may be guided by different scenario-influenced needs. For example, one may try to look for more popular datasets to connect, where popularity is defined as a function of the number of links received by other datasets. By doing so, the visibility of the source dataset may increase. Toupikov et al. (2009) defined a dataset ranking algorithm using VoiD dataset descriptions, that can be used for this purpose of identifying the most popular datasets. One can instead fix the target dataset and try to make the set of links more *complete*. The notion of completeness has been seen from different points of view.

Some authors interpreted that source entities should link with cardinality 1:1 to entities in the target dataset (Zaveri et al., 2016). Albertoni and Pérez (2013) proposed a set of measures that assess the extent to which the set of links cover the concepts and entities in the connected datasets, and signals what unconnected concepts and entities can be further linked. (Beek et al., 2018) applied equivalence closure to existing `owl:sameAs` links based on “reflexivity, symmetry and transitivity" to derive new links between the entities at hand. Both approaches focus on “completing" the link sets, with the given set of entities, classes and properties.

In their book, Heath and Bizer (2011) present a list of recommendations on how to create and publish Linked Data. When it comes to assessing potential datasets for linking a dataset, the authors recommend to answer questions such as “what is the value of the data in the target dataset?", “to what extent does this add value to the new dataset?" and “are there ongoing links to other dataset so that applications can tap into a network of interconnected data sources?". By measuring completeness and accuracy as explained above, it is not possible to provide an answer to the aforementioned questions on the value added by the interlinking. Hence, there is a need to investigate new dimensions and measures that aid data publishers assess the information gain their entities have via the interlinking, such that they can improve the perceived value by adding new links and/or updating the links they created. This assessment can be determined at an entity, link or dataset level.

1.3.3 Managing Crowd/Community Contributions

Due to the novelty and complexity of these hybrid human-machine systems, there are still many open problems, ranging from human behavior prediction to human-machine coordination. Here, we concentrate on two challenges that relate to analyzing and recognizing the crowd’s contributions.

1.3.3.1 Contribution Inequality and Attrition

Any kind of hybrid (human-machine) system needs to sustain human input to be able to well-function. While there are potentially millions of humans connected to a myriad of devices every day, who could provide input to a human computation system, there is a phenomenon that tends to reoccur in many social systems: substantial contributors are scarce. Even in large communities, there is usually an imbalance between active and passive usage, which distinguishes those who contribute from those who read. Ideally, these hybrid systems should aim to have a large volume of diverse people contributing repeatedly. Volume is necessary, because typically the human computation “work" to be done is abundant, especially in times when we have large datasets to be processed. Diversity is important, because people with different backgrounds, skills and interests can accomplish a wider variety of tasks and curate richer knowledge. In terms of turnover, it is better to have people working on a type of task/system repeatedly for a longer period of time, because on the one hand, people tend to master tasks over time, and on the other hand, task switching can add a cognitive load on people that diminishes the quality and/or efficiency of their work (Rubinstein et al., 2001; Difallah et al., 2016). Additionally, the process for onboarding new users to make them get used to workflows, technical details and social interactions can require the investment of many resources.

In order to achieve these ideal characteristics, strategies to *attract*, *engage* and *retain* users need to be specified and implemented within systems. Designing methods that succeed in accomplishing these three actions effectively is very difficult, because they all entail some kind of human behavioral change. Furthermore, the application of these methods can have undesired results when they are not tailored to the needs of the scenario at hand (e. g. engaging malicious users would lead to detrimental results).

Several platforms combining human and machine computation with the help of large crowds have been designed under different incentives schemes, which in one way or another influence the methods used to gain and keep people working on tasks. In the context of microtask crowdsourcing, requesters offering crowd work sometimes post the announcement about their freshly published microtasks in social media (e. g. in Twitter), and workers often post recommendations on available crowd tasks in dedicated fora. With the purpose of attracting workers with more specific technical knowledge, Ipeirotis and Gabrilovich (2014) created a method that strategically placed targeted advertising in sites and invited e. g. people with medical knowledge to participate in related crowd tasks in Amazon Mechanical Turk. The intervention to increase user engagement is usually implemented at a task-level (Law and Ahn, 2011), so every requester has the freedom and responsibility to design their. Besides maximizing the number of microtasks in a batch and the economic reward (Difallah et al., 2014), requesters can experiment with tweaking their tasks toward human factors that influence user engagement, such as feeling entertained (Elmalech et al., 2016) or curious (Law et al., 2016). Retention is by contrast not that much in the hands of requesters, unless they establish additional measures, such as linking to further tasks they publish or manually contacting workers who participate in their tasks to forward them to further tasks. Wikidata, the free and multilingual knowledge base that anyone can use and edit, has a very different way of operating: Wikidata contributors are intrinsically motivated to participate. As a peer-production system (Benkler et al., 2015), Wikidata is curated and maintained by a community of volunteer contributors, and the automated means they develop (i. e. tools and bots). The reward for their contribution is not economic. However, there is community recognition and accountability, which tends to motivate large amounts of people to deliver high quality work diligently. Potential Wikidata contributors are attracted by some of the sister projects' popularity among Web users (e. g. Wikipedia), via specific community events (e. g. editathons or hackathons), as well as scientific activities of disciplines interested in its technological pillars (e. g. Semantic Web and Computer-Supported Cooperative Work (CSCW) research). Engagement management is typically operationalized via social interaction (e. g. discussions, messages of gratitude) or the proliferation of thematic sub-projects that try to direct contributors to their own interests. Wikimedia projects, including Wikidata and Wikipedia, automate email deliveries that remind and recognize users for doing their n-th round contribution.

All in all, retention management is one of the hardest problems in crowd contributions' management, and it remains unsolved. At the same time, providing solutions that help reduce user attrition is imperative for ecosystems like Wikidata, where in principle, every edit done with good faith counts. While the (scientific) community has investigated editing patterns of Wikipedia editors extensively, in terms of collaboration (Walk et al., 2016) and evolution (Panciera et al., 2009), analogous studies analyzing Wikidata contributions are scarce. More specifically, there is no data-driven study focusing on the differences between editors that contribute long and/or to a large extent, and editors who drop after a while/some edits. Past research has also looked into increasing retention of new comers (Ciampaglia and Taraborelli, 2015) and the detrimental impact of reverting new comers' edits on their engagement and the quality of their work (Halfaker et al., 2011) in the context of Wikipedia. While Wikidata (and any other Wikimedia project) has methodological similarities with Wikipedia regarding community organization and peer-production, the many more options for contribution and the intrinsic structured nature of its content could potentially influence editing behavior. Therefore, there is a clear need to investigate methods that inform community managers about editor attrition in Wikidata, so that they can improve their retention strategies.

1.3.3.2 Cross-Platform Crowd Work Recognition

A different problem present in microtask crowdsourcing platforms, that is often overlooked but places crowd workers in the middle of a power imbalance situation, is the lack of cross-platform work recognition. Crowd workers can contribute to any of the hundreds of marketplaces, but every time they sign up for a new one, they are considered novices and their actual skills, knowledge and previous experience is invisible to requesters. Such a lack of information can influence negatively the crowd workers' options to be selected for tasks with specific requirements.

Current platforms, including Figure-Eight and Amazon Mechanical Turk, maintain a history of worker assessments as workers accomplish more and more microtasks. This information remains locked within the platforms. To identify competences of workers and filter unsuitable workers, requesters often create ad hoc qualification tests within their task batches. Several scientific works on task assignment and task scheduling looked into profiling techniques, that tried to learn expertise (Khazankin et al., 2011a; Satzger et al., 2013), interests (Difallah et al., 2013a) and even personality traits (Kazai et al., 2011) from crowd workers in order to match tasks and most suitable workers, with the goal of increasing the output's quality. However, the majority of these methods operate at a task-level.

In order to improve the situation in terms of transparency, fairness, and worker pre-selection efficiency, there is a need to create methods that facilitate the exchange of work registries among workers, platforms and requesters.

1.4 Summary of Scientific Contributions

In this section, the major contributions of this thesis are summarized. Each of the contributions tackles one of the research challenges listed above. [12]

1.4.1 A Feasibility Study on the Use of Crowdsourcing for Ontology Alignment

Having Challenge 1.3.1 in mind, we investigated a crowdsourcing-based approach to aligning ontologies in the Semantic Web. The main goal of this work was to *explore whether ontology alignment is amenable to microtask crowdsourcing, as well as to identify design constructs that increase the effectiveness of our approach*. Our three concrete research questions and corresponding findings are summarized below.

RQ1: Is ontology alignment amenable to microtask crowdsourcing?

To answer this question, we defined and implemented CrowdMap, a solution that given a set of mappings produced by an ontology alignment algorithm and the original ontologies aligned, automatically generates and publishes microtasks that are deployed in the crowdsourcing platform CrowdFlower –nowadays called Figure-Eight, retrieves the judgment of multiple crowd workers per mapping and computes a final alignment based on the filtering and improvement of the mappings by the crowd. We selected this type of crowdsourcing (as opposed to others targeting freelance experts) because we needed a scalable and flexible method to be able to process potentially large ontologies of the Web of Data. From all the available options for publishing crowdsourcing microtasks, we decided to use CrowdFlower because as a marketplace aggregator, it distributes the tasks to multiple marketplaces

[12] These summaries are the result of adapting and synthesizing the text included in the core contributions' chapters in this thesis.

(including, at the time of the experiments, Amazon Mechanical Turk). Hence, with CrowdFlower we maximized the potential size (and possibly diversity) of the pool of crowd workers that provide human computation in our scenario.

While the approach can easily be extended and applied to further types of data, in our study we concentrated on the mapping of classes, and the set of mapping relationships including *equivalence* (=) and *subsumption* (<=, >=). CrowdMap uses the CrowdFlower API to materialize the UI and workflow of the microtasks, so as to collect the individual answers of crowd workers per reviewed mapping and published microtask. The platform provides basic form controls and features to publish groups of microtasks that share the same UI and configuration (i. e. number of workers per microtask, the time workers may invest while solving the task, minimum accuracy that workers must demonstrate in an initial test phase). Crowd workers provide their input individually, without collaborating, and (at the time of the experiment [13]) workers can only be filtered by country of origin.

We designed ontology alignment microtasks of two types: *validation* and *identification*. In the former, crowd workers indicate whether the mapping is valid or not, while in the latter, they have to select the type of relation that exists between the pairs of classes. Besides the standard elements including title, instructions, problem statement and form to collect the answers, our microtasks contained so-called verification questions that we implemented in order to identify and avoid malicious crowd workers, who click randomly/systematically on the same answers across tasks. In the problem statement, we presented crowd workers with the labels of the classes and their definition, which we obtained automatically from the ontology implementation. We additionally published tasks in two halves, in order to shuffle the order of the options in the forms. Moreover, we followed CrowdFlower's system for quality control based on so-called "golden units"—microtasks that contain the ground truth and that are used to test the accuracy and consequently trustworthiness of crowd workers in that specific group of tasks. Further details of the microtasks can be inspected in Chapter 3.

We conducted an empirical evaluation using pairs of ontologies and alignments produced by automatic algorithms from the Ontology Alignment Evaluation Initiative (OAEI) –a scientific benchmarking campaign co-located with the International Semantic Web Conference for over a decade now. We selected ontologies from two tracks from the campaign and used precision and recall as evaluation measures, following the methodology defined at the OAEI.

After running several pilots where we generated different versions of the microtasks, we concluded that there were several factors that influenced the quality of the results: First, deploying the microtasks without any kind of quality assurance mechanism led to fast but low quality results. Second, the selection of gold units (used to test and filter out crowd workers by proven task-specific accuracy) is crucial; they should show a variety of cases in order for the workers to learn how the microtask should be done, and they should contain no ambiguity. Third, we observed that the wording of the question in such an abstract matter has impact on the final results, as crowd workers seemed to understand better the question "Is Concept A the same as Concept B?" than "Is Concept A similar to Concept B?". Fourth, the approach relied on the documentation of the ontology, as labels and concept definitions make the readability of the ontology better, and the meaning of concepts clearer.

Once we obtained the most effective formulation from the initial pilots, we ran experiments with different data configurations, publishing the microtasks in CrowdFlower and collecting responses from real crowd workers from different countries, including the US and India. In order to answer this research question, we generated a set of mappings by computing the Cartesian product among the classes

[13] At a later stage CrowdFlower added the possibility to select from three possible groups of people, that range from least accurate and fastest option to most accurate and slowest option.

of the source and the target ontologies whose alignment we used to evaluate our approach. The resulting alignment by the crowd, better than random, led to perfect recall but weak precision (0.53). Hence, the approach helped in finding all the mappings that the classes of the two ontologies should have, but at the same time introduced quite some noise (with many false positives).

RQ2: How does such a human-driven approach compare with automatic (or semi-automatic) methods and techniques, and can it improve their results?

To find an answer to this question we ran experiments with two more configurations. On the one hand, we computed the crowd results for the candidate alignment produced by one of the state-of-the-art algorithms (i. e. AROMA algorithm, which had had a strong performance in the 2011 OAEI edition). On the other hand, we generated a synthetic set of mappings, simulating 1.0 recall and 0.50 precision, which is indeed very close to the results obtained by the Cartesian product setting. Our findings showed that the the average CrowdMap performance was better than the compared algorithms, and we saw that the crowd is able to improve precision. Given the fact that the CrowdMap approach is able to obtain such a good values for recall, we foresee that a two-step process could be implemented, following the Find-Fix-Verify scheme proposed by Bernstein et al. (2010), where in a first round we would aim to find maximum recall, and in a second round we would try to clean the set of false negatives to improve precision.

RQ3: What types of alignment problems can workers feasibly solve? What correspondences between elements of different ontologies (e.g., similar, more general, more specific) can be reliably identified via crowdsourcing?

Our experiments included ontologies with all three kinds of relationships, and the two aforementioned types of tasks (mapping validation and mapping identification). In our experience, and with the tested ontologies, there was no remarkable difference and crowd workers were able to provide correct answers in a similar way, in any of these variations. The differences were appreciated when crowdsourcing the alignment of different ontologies, probably due to the nature, modeling and degree of documentation of the ontologies.

In summary, with this study we proved that it is feasible to use microtask crowdsourcing for the task of ontology alignment, assuming the ontologies have similar characteristics to ours in terms of domain complexity and readability. CrowdMap provides a solution to augment automatic methods for ontology alignment in a cost-efficient, fast, and scalable manner.

This first investigation paved the way for further related works, which used crowdsourcing to curate ground truths for evaluation efforts, or improve the results of ontology alignment algorithms either at creation or publication time.

1.4.2 Information Gain-based Measures to Intrinsically Assess Link Quality in the Web of Data

To address Challenge 1.3.2, we defined, implemented, and evaluated a set of measures that assess existing links connecting two datasets in terms of what they add to the source dataset (i. e. the dataset the links originate from), in a context-independent manner. The overarching goal of this research work is to *investigate how to measure the information gain enabled by links, in order to inform data publishers/managers on potential points of improvement in their interlinking*. We envision data publishers

making use of this information, both at design and reviewing time, when they have either a sample or the entire set of links.

We first identified the key principles for data interlinking based on foundations of the field of Information Retrieval. We considered basic query needs that users of a holistic Web of Data typically have: to retrieve the information with highest precision and recall possible, be able to use different terms when querying and still find answers, and understand the relation between existing data. The identified principles suggest the linking to be designed such that:

P1 it extends the *description of entities* of the source dataset.

P2 it increases the number of *entities and datasets that source entities are connected* to.

P3 it makes the source entities have a description with a higher number of *vocabularies in their description*.

Based on these principles, we defined a set of measures that capture the extent to which a set of links impact these three dimensions: entity description, entity connectivity and use of various vocabularies. In order to determine **gain** we compare the measurements of two different states of the source data, namely the data without the external links (`state_1`), and the data including the external links (`state_2`) [14]. There are various ways to implement gain measures: a straightforward possibility consists of counting the elements that are present in `state_2` and not in `state_1`, where elements refer to either distinct entity description statements, entities or vocabularies. By doing so, gain refers to the number of *new* things discovered when considering links. If no new things are discovered, then the gain is equal to zero. While this measurement may already be informative, it does not signal redundancy, which can be relevant for data managers who would like to be aware of (a) duplicated facts distributed throughout several locations to be able to update them on cascade, or (b) the "data overload" the entity descriptions have in their datasets in terms of, for example, predicates. The Shannon Entropy (Shannon, 2001) has been used in the literature to capture diversity, and consequently redundancy, as one is the inverse of the other. We defined our measures using the diversity index and the concept of information gain characterized by the Shannon Entropy as a basis, and specifying views of the data for each of the three principled dimensions (entity description, entity connectivity and use of various vocabularies). Our measures are intended to provide comparative assessment instead of an absolute assessment. Hence, they indicate which entities gain more or less with the links, considering any kind of link. The person or application inspecting the measurements should interpret its meaning, and make a decision accordingly (e. g. a data publisher eager to improve the interlinking can decide where to start from by focusing on entities with lower gain).

In order to empirically evaluate our measures, we used real-world data from the Linked Data Crawl dataset [15]. We first tested the validity of the measures looking at requirements that measures should meet in general, and we then inspected the measurements obtained for the datasets' entities to check if (i) redundancy exists, and (ii) datasets show an heterogeneous distribution of information gain.

RQ4: Are our diversity-based measures *valid*?

[14] Note that one could implement the same process to analyze the gain that a linking in version v2 (`state_2`) has to offer in comparison to the linking in version v1 (`state_1`)

[15] Snapshot of the LOD in 2014 crawled to analyze its status `http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/`

Following the methodology specified by Behkamal et al. (2014) to validate linked data quality measures, we computed measurements for the 6 measures that we defined and 35 datasets selected from the Linked Data Crawl. In our analysis we tested whether the measures are discriminating independent.

Discriminating measures (i. e. they produce distinct values for different entities and datasets.) We analyzed the mean, standard deviation and quartiles of all measurements for all datasets. Looking at the obtained data files, we confirmed the hypothesis, as the values for the measures vary across datasets, except for the classification extension, where all datasets show a mean, standard deviation and quartiles of 0.0 for the difference in entropy.

Independent measures (i. e. their values are not inter-correlated.) We computed the Spearman correlation of all the measurements within each dataset, including all types of links. The results indicate that the two measures looking at connectivity are highly correlated, which makes sense since one examines target entities and the other target datasets. With these results we determined that having only one of the two for measuring connectivity gain (Principle P2) is sufficient. The rest of the measures are independent.

All in all, we conclude that in the context of our evaluation data, four out of our six measures are fully valid (description extension, predicate description extension, entity connectivity extension and increase in the number of vocabularies used), and one (classification gain) is partly valid (i. e. it is independent but not discriminating).

RQ5: Is it meaningful to obtain diversity-based gain measurements from current Linked Data?

We envision the target user of these measurements, either analyzing a sample of the links to adjust the interlinking configuration, or reviewing a fully executed interlinking with the idea of adding new types of links or new target datasets. In either of such cases, the requirement for our measurements to become meaningful is that they are able to highlight the properties they are singular for: they highlight entities with more/less information gain, as well as redundant information. We explored the values obtained by our measures in the real-world LOD datasets of our evaluation, and focused in the following aspects:

Entities in datasets show inequality in terms of the multi-dimension information gain. We studied how distinct entities of the same dataset gain information to different extents. By visualizing the computed measurements per dataset in boxplots, we were able to examine such inequality: this visual representation allowed us to examine groups of entities sharing the same range of values, the distance between the minimum and the maximum values in the whiskers (the boxes), as well as the existence of outliers (cases that notably differs from other observed cases (Grubbs, 1969)). We thus generated boxplots showing the measurements per dataset, per measure and per type of link, distinguishing between identity links (e. g. owl:sameAs), relationship links (e. g. wgs84:location), classification links (e. g. rdf:type), similarity links (e. g. skos:closeMatch), and other more general links (e. g. rdfs:seeAlso). We observed these in terms of classification, description, connectivity and vocabulary heterogeneity –the major dimensions that our measures adhere to. From the results (explained in detail in Chapter 4), it can be highlighted that most of the datasets showed outliers; hence, there is the opportunity to inspect and improve these individual cases. Note that such outliers could be either entities that do not gain as much information as other entities, or entities that gain an extraordinary amount of information. In the latter case, the gain could be genuine, or due to an inaccurate link that is aligning

entities on identity, when in reality are not referring to the same thing. The length of the whiskers in some of the measurements was of several units (e. g. for description extension in identity links, and entity connectivity in identity links), which suggests that there are several groups of entities gaining differently. The median across datasets varies, suggesting that data publishers could also use that as for some sort of *social comparison* to aim for gaining as much as others do. Classification is the dimension that gained the least, while description extension showed the highest gain. The vast majority of datasets gained one additional vocabulary with the linking.

Current LOD datasets have entities with zero or negative information gain. In terms of redundancy, we observed that for all types of links there are datasets that have entities with a negative gain in predicate-based description extension. This means that in the resulting linked data there are entities that contain multiple statements using the same predicate. Having such predicate-redundant statements might be due to the fact that these predicates are not functional properties, or because statements contain noisy/erroneous values. Either way, adding statements of the same predicate does not add diversity to the description.

With these results we concluded that the measures we propose fulfill the requirements necessary to show potentially meaningful observations for improving interlinking in terms of diversity.

1.4.3 Analysis and Predicting the Evolution of Editors in Wikidata

Concentrating on the human side of human-machine systems, and with the purpose of providing Wikidata community members valuable observations (Challenge 1.3.3.1), and methods to understand and intervene in the community, we conducted a longitudinal analysis of the Wikidata edit history. The overarching goal of this work was to *study if there are differences between so-called power and standard editors in terms of editing patterns over time, that help predict if an editor will or not become an engaged editor.*

Marketing research has a long tradition of analyzing customers and implementing churn management (i. e. the process of identifying, understanding and trying to reduce the loss of customers) (Rosenberg and Czepiel, 1984; Ang and Buttle, 2006). Studies have shown evidence that targeting customers can have positive impact to improve retention (Verhoef, 2003). However, customers need to be approached in a tailored way, because loyal customers need a different attention than likely-to-drop customers. A typical prediction task in churn management is the prediction of whether a customer will be a *churner* or not (Vafeiadis et al., 2015). Once this information is predicted, churn managers can intervene in order to convince potential churners to change their behavior. In order to design this intervention carefully, it is useful to have an indication of the time when the customers' engagement will decay, to be able to approach them before that happens or when it just starts to happen. Even though the reason to keep editors contribution in the Wikidata ecosystem is not economic, there is attrition and participation inequality. Hence, gaining a deep empirical understanding of the actual situation and using assumptions and techniques from this field of research becomes valuable.

In this work, we were precisely interested in making predictions in terms of (i) the time that contributors will be in Wikidata (i. e. their *lifespan*), as well as (ii) the *volume of edits* they will contribute with, to be able to understand the magnitude of their future engagement. Our two major research questions were:

RQ6: What is the distribution of contributions in Wikidata?

RQ7: Are there differences in editing behaviour between power and standard editors that help predict engagement?

In order to provide an answer to these questions, we first carried out an exploratory longitudinal analysis to understand existing differences between editors who show different engagement levels – so-called power and standard editors. Secondly, we provided and evaluated a predictive model, using insights from the exploratory data analysis.

Since the goal was to measure lifespan and volume of edits, and each of these two variables can have 2 possible values, contributors can be classified in 4 types of contributors:

1. contributors with **long lifespan** and **high volume** of edits
2. contributors with **short lifespan** and **high volume** of edits
3. contributors with **long lifespan** and **low volume** of edits
4. contributors with **short lifespan** and **low volume** of edits

The contributors with the highest impact on the system are contributors who contribute extensively, and for a long time (group (1)). Groups (2) and (3) are also valuable contributors. For example, a contributor supervising recent changes to revert and correct malicious edits, a few times a month; (s)he might do only a couple of edits a month, but if she does it for a long period of time, her contribution can help Wikidata in terms of data quality. Yet, ideally one would like to have as many contributors as possible in group (1). In the following, we refer to power editors as the editors who present a large contribution (in terms of either lifespan or volume of edits), and standard editors as the editors whose contribution is limited (again, in terms of either lifespan or volume of edits).

Predicting the lifespan and volume of the contribution of editors, we are able to classify existing contributors into one of these groups, and we could in the future, decide consequently whom to address and how to do it. It is worth mentioning that implementing the intervention to convert editors from one of these categories to another was not the scope of this work, as that would be a different contribution.

In order to make such a prediction, we analyzed the evolution of editing behavior. We analyzed this information from two different perspectives: we ran a (i) *session-based* analysis, and we also study the editing progress (ii) on a *monthly basis*. The two perspectives are complementary: with the former, we aimed at understanding the extent to which editors change their behavior as they gain more experience and do more edits in each session they spend editing; while with the latter, we performed a time-sensitive analysis.

When we analyzed the behavioral evolution throughout the editors' lifetime (in sessions and in months) we measured *indicators related to the editors' productivity, editors' participation and the diversity of the types of edits*. We decided on these three dimensions, based on the following findings published in the related work:

1. The seminal work by Panciera et al. (2009) revealed that "Wikipedians are born, not made", and showed that they maintain a constant level of contribution. Hence, we hypothesized that a constant contribution over time is a signal of power editors but not of standard editors.
2. Editors who develop editing habits, are likely to schedule them regularly in their agendas, and the longer a habit runs for, the more established it becomes (Duhigg, 2012). This argument suggests that a constant participation over time is a signal of power editors but not of standard editors.

3. Piscopo et al. (2016) surveyed Wikidata editors to find how their tasks change while they move from novice to proficient editors. In their qualitative study, the authors discovered that editors take more responsibility and do different tasks over time. With these findings we hypothesized that an increasing trend in the diversity of the tasks over time is a signal of power users but not of standard users.

In order to test these hypotheses, we analyzed the edit history of Wikidata over the course of 4 years. The Wikidata dumps published by Wikimedia contain the set of all edits, annotated with the author of the edit –that be an anonymous IP, a registered human user or a registered bot user, the timestamp when the edit was implemented, the item/project page where the edit took place and an optional comment. After pre-processing the data and generating edit sessions, computing the time between edits and following the methodology proposed by Geiger and Halfaker (2013) in Wikipedia, we obtained descriptive statistics in terms of volume of edits and lifespan. The major findings of this initial analysis were the following:

1. There is a skewed distribution of edit counts (i. e. few editors with many edits and vice versa).

2. There is a skewed distribution of editors per item (i. e. few items are edited by many editors and vice versa).

3. There is a slightly skewed distribution of lifespan (i. e. few editors were in the project for many months and vice versa).

4. There is not a linear relation between the lifespan and the volume of edits done by editors.

5. Compared to Wikipedia, there are many more edits that occur in a shorter timespan (of a few seconds). One the one hand, the edit granularity of Wikidata is much smaller. On the other hand, it can be that when editing data facts, users keep open multiple tabs and edit different statements/items consecutively.

6. We empirically define new sessions after 4.37 hours of inter-edit time, around 4 times longer than in Wikipedia.

As a next step, we computed a set of indicators to measure the editors' contribution, participation and the diversity of the types of edits over time. We specifically measured the number of edits (i1), the average number of edits per item (i2), the number of items edited (i3) for contribution, the number of seconds dedicated per session (i4) for participation and diversity of types of edits (i5) for diversity. We obtained the two temporal views of the data (i. e. measured indicators i1-i5 per session, and measured indicators i1-i3 and i5 per month), and observed how editors evolved according to the indicators. In order to study the evolution, we fitted a linear model using the RANSAC algorithm (Fischler and Bolles, 1981), which gave us the shape of the evolution in terms of the slope and intercept over time. We compared the evolution of the two different parts of the population –standard and power edits– which we delineated empirically in terms of volume of edits and lifespan. This analysis led to the following results:

1. Editors with long lifespan have a constant contribution over months, while editors with short lifespan do not.

2. Editors with a high volume of edits have a constant participation over sessions, while editors with low level of volume of edits do not.

3. Editors with a long lifespan tend to increase the diversity of the type of their edits, while editors with short lifespan can either increase or decrease it over the months.

After this quantitative analysis, we reached the following conclusions with regard to the initial hypotheses: the difference between editors with long / short lifespan and editors with high / low volume of edits, in terms of the way they evolve is in some cases obvious and in others not. We can confirm Hypothesis 1, because we see that editors with high lifespan show a constant contribution, while other do not. The same applies to high volume of edits, although with less strength (cf. Finding 4). We can also confirm Hypothesis 2 in the case of lifespan, because we see that people with longer lifespan maintain a constant participation, while editors with shorter lifespan do not. However, when it comes to volume of edits, the measurements do not help differentiating the two groups of editors as clearly. Hypothesis 3 is rejected, because we see that both in long and short lifespan, and in high and low volume, editors tend to increase the diversity of the type of actions they accomplish. Hence, evolution in contribution and evolution in participation are good indicators to differentiate standard and power editors in (especially in terms of lifespan), while diversity alone it is not sufficient.

As a last step, we defined supervised classification methods to predict the range of months that an editor will be contributing to Wikidata, and the range of edits that an editor will implement in Wikidata. In other words, given an editor, and a set of observation based on the aforementioned productivity, participation and diversity indicators (i1-i5), we defined two prediction tasks:

- whether the editor will be contributing with a high or low volume of edits
- whether the editor will contribute for a long or short lifespan

We implemented two classifiers (logistic classifier and random forest classifier) and evaluated the prediction based on F1-score. Despite the unbalanced nature of the data (i. e. we had few editors with many edits or long lifespan), we observed that it is possible to automatically predict the future the volume of edits and lifespan duration of an editor based on the available edit history of Wikidata editors. We are able to obtain better prediction results than a naive classifier in each of the 4 tested configurations. We were able to predict lifespan better than volume of edits (with an average F1 score above 0.9) in both session- and month-based evolution. Between the two prediction problems that we set up (i. e. predicting the lifespan range and predicting the range of volume of edits), predicting the lifespan has a higher priority, because it gives us the key information about when we should address standard editors that will become inactive. Therefore, being able to predict lifespan more accurately than the volume of edits is a positive result.

To close the contribution, we highlighted a set of implications that these findings could have in terms of community management design.

1.4.4 An RDF vocabulary to model crowd worker activity and encourage cross-platform work recognition

In order to contribute to Challenge 1.3.3.2 and facilitate the exchange of crowd work registries, and better recognition for crowd work, we designed and implemented an ontology for crowd work curricula vitae (CVs). The main goal of this work was to propose and discuss with the crowdsourcing community,

the information needs that should be covered in such a CV scenario. Nonetheless, it is worth mentioning that the implementation of a real system that allows to extend crowdsourcing platforms with Crowd Work CV features was out of the scope of this thesis.

As a first step, we reviewed the information that multiple crowdsourcing platforms collect and share about crowd workers, and we identified a set of requirements that such an ontology or vocabulary must fulfill:

- it should be domain independent to maximize reusability of crowd work activity reports.
- it should be defined avoiding marketplace-specific rules.
- it should be syntactically and semantically interoperable.
- it should be easily extensible.
- it should comply with standard (de facto) CV systems such as Europass or LinkedIn.

With these requirements in mind, we opted to use Semantic Web technologies, which facilitate the definition of vocabularies with precisely these characteristics. The Crowd Work CV ontology that we specified in OWL describes crowdsourcing agents (i. e. crowd workers and requesters), their interests, obtained qualifications and work history. We followed standard ontology engineering practices and reused related ontologies, including FOAF [16] and SIOC [17]. We also aligned our crowdsourcing-oriented ontology, with other more general-purpose ontologies such as ResumeRDF [18]. The ontology has classes and properties to describe user accounts, qualifications, microtasks, marketplaces and assessments that crowd workers receive after accomplishing crowd tasks.

We validated the quality of the ontology with a well-known ontology engineering verification tool and methodology, that considers a list of 40 common pitfalls in ontology specifications. Additionally, we published a survey in CrowdFlower and collected 200 responses from crowd workers to check whether crowd workers would be willing to adopt a Crowd Work CV-based systems. The results indicate that crowd workers would generally be interested in reporting and sharing this kind of information to make their cross-platform work accountable during worker to task assignments (see Chapter 6 for further details).

1.5 Overview of this Thesis

This thesis is organized as follows: This first chapter is the introduction, where relevant topics are motivated, the major challenges addressed in this this thesis are introduced, and the scientific contributions of the thesis are summarized. Chapter 2 provides an overview of foundational concepts that are essential for the understanding of the rest of the thesis. Chapters 3 to 6 are the core scientific chapters that include each of the aforementioned contributions. Chapter 7 wraps this thesis with a set of concluding thoughts and an outlook to future challenges for the field of human-machine link quality management and related scientific areas.

The contributions of this thesis have led to the following scientific publications:

[16] FOAF vocabulary `http://xmlns.com/foaf/spec/`
[17] SIOC ontology `https://www.w3.org/Submission/sioc-spec/`
[18] ResumeRDF `http://rdfs.org/resume-rdf/`

Chapter 3 This chapter includes the scientific publication containing the contribution introduced in Section 1.4.1, to address the challenge presented in Section 1.3.1.

Cristina Sarasua, Elena Simperl, and Natalya F. Noy. (2012). **CrowdMap: Crowdsourcing Ontology Alignment with Microtasks**. In Proceedings of the 11th *International Conference on The Semantic Web* - Volume Part I (ISWC2012), Philippe Cudré-Mauroux, Jeff Heflin, Evren Sirin, Tania Tudorache, and Jérôme Euzenat (Eds.), Vol. Part I. Springer-Verlag, Berlin, Heidelberg, 525-541.

The initial idea of using microtask crowdsourcing for the task ontology alignment was proposed by Elena Simperl and Natasha Noy, as a first step towards using CrowdFlower/MTurk crowd workers for a Semantic Web ontology task. Cristina Sarasua designed and implemented both the solution and the empirical evaluation, and wrote the content of the paper. Elena Simperl and Natasha Noy provided feedback throughout the experiments and the writing phase of this work.

Chapter 4 This chapter includes the scientific publication containing the contribution introduced in Section 1.4.2, to address the challenge presented in Section 1.3.2.

Cristina Sarasua, Steffen Staab, Matthias Thimm. (2017). **Methods for Intrinsic Evaluation of Links in the Web of Data**. In Proceedings of the 14th *Extended Semantic Web Conference* (ESWC 2017), Eva Blomqvist, Diana Maynard, Aldo Gangemi, Rinke Hoekstra, Pascal Hitzler, and Olaf Hartig (ed.), The Semantic Web - Portorož, Slovenia, May 28 - June 1, 2017, Proceedings, Part I., Vol. 10249 (pp. 68-84). Springer, Berlin, Heidelberg.

The conceptualization of the idea, the design and implementation of both the solution and empirical evaluation of this work was done by Cristina Sarasua. Cristina Sarasua collected and prepared the data, implemented the code and ran the experiments independently. She also wrote the complete paper, and incorporated the feedback received from the senior co-authors, especially Steffen Staab.

Chapter 5 This chapter includes the scientific publication containing the contribution introduced in Section 1.4.3, to address the challenge presented in Section 1.3.3.

Cristina Sarasua, Alessandro Checco, Gianluca Demartini, Djellel Difallah, Michael Feldman and Lydia Pintscher. (2019). **The Evolution of Power and Standard Wikidata Editors: Comparing Editing Behavior over Time to Predict Lifespan and Volume of Edits**. *Journal of Computer Supported Cooperative Work (CSCW)* Springer p. 1-40

Cristina Sarasua devised the idea and design of the paper. She took care of the complete large-scale data preparation, including processing the Wikidata edit history, implementing and extracting the editing behavior features. She also worked on the data analysis and the implementation of the predictive models. For the data analysis, Cristina Sarasua worked in tight collaboration with Alessandro Checco, who as a more experienced data scientist and applied statistician, provided very valuable feedback and helped in the implementation of tasks such as the definition of sessions via statistical methods, and the measurement of editing behavior evolution. Cristina Sarasua wrote most of the content of the journal article and implemented the feedback of the senior co-authors, especially Alessandro Checco and Gianluca Demartini. Djellel Difallah assisted in database administration tasks, helping with machine performance issues in a specific server hosted by Djellel Difallah's organization, where Cristina Sarasua executed the queries that she implemented.

Chapter 6 This chapter includes the scientific publication containing the contribution introduced in Section 1.4.4, to address the challenge presented in Section 1.3.3.

> Cristina Sarasua, Matthias Thimm. (2014). **Crowd Work CV: Recognition for Micro Work**. In Proceedings *Social Informatics* (SocInfo 2014) Workshops: 429-437.

The conceptualization of the idea, the design and implementation of both the solution and testing of this work was done by Cristina Sarasua. Cristina Sarasua implemented and documented the ontology. She also wrote the complete paper, and incorporated the feedback received from Matthias Thimm.

Other Revelant Publications

Additionally, the author of this thesis was the first author of a viewpoint article "defining a roadmap for the field of Crowdsourcing and the Semantic Web", that resulted from group discussions held in a *Perspectives Dagstuhl Seminar* on crowdsourcing and the Semantic Web. The article published at the Human Computation Journal has been excluded from this manuscript; however, given its relevance to the topic of this thesis, it is worth citing it here.

> Cristina Sarasua, Elena Simperl, Natalya F. Noy, Abraham Bernstein, Jan Marco Leimeister. (2015). **Crowdsourcing and the Semantic Web: A Research Manifesto**. *Journal of Human Computation*, 2, 3-17.

The complete list of publications by the author of this thesis can be found in the CV attached.

Chapter 2

Foundations

In this chapter, I introduce fundamental concepts related to the technology used throughout the work presented in this thesis. First, I describe Semantic Web technologies. Specifically, in this part RDF graphs and ontologies are defined, the tasks of ontology alignment and link discovery are specified, and existing algorithms for such tasks are described. Additionally, a well-accepted multi-dimensional definition of data quality is presented and its relation to Semantic Web data is summarized. Second, I introduce Wikidata, which is a unique RDF graph, created, linked and maintained by a combination of humans and machines, and therefore, a very relevant case study for the problems investigated in this thesis. I describe the basic characteristics of Wikidata's data, as well as the peer-production process around it.

2.1 Modeling, Publishing and Consuming Semantic Web data

The Semantic Web vision, as initially explained in a seminal article by Berners-Lee et al. (2001), proposed the annotation of Web documents with machine-readable metadata, paving the way to intelligent agents that would traverse these Web documents, comprehending their meaning and inferring facts from them, to be able to assist humans in solving tasks such as search or event scheduling effectively. Nowadays, private and open parts of the Web are densely populated with such machine-readable annotations. While the more visionary autonomous agents scenario is not universally in practice, there are many information retrieval and data management systems using the original Semantic Web technologies (e. g. RDF, RDFa, RDFS, OWL, SPARQL) or derivatives of them (e. g. Schema.org, JSON-LD), solving tasks such as Web search and automatic scheduling. To name a few examples of products reaching millions of users with technology based on RDF, Google's search engine adopted semantic search, Google Calendar can automatically add events extracted from emails received at Gmail thanks to the RDF-based annotations serialized in JSON-LD, and Airbnb developed their own RDF graph to help categorize their data and provide better travel context to users.

These technologies allow to express the meaning of things explicitly and unambiguously via metadata, increasing the interoperability between systems. In order to do so, the knowledge representation techniques used in the Semantic Web (with foundations in *Description Logics* (DL)) (Antoniou et al., 2012)) differentiate between the *ontology*, which models characteristics of types of things, and the *data*, which describes characteristics of concrete instances using the ontology constructs. Both objects (schema and data) can be queried separately.

Throughout the years, the value of Semantic Web technologies for Web data integration was widely acknowledged, even in cases where the capability of logics to generate new knowledge through inference was scarcely used. After Tim Berners-Lee proposed what he coined *Linked Data* (Shadbolt et al., 2006; Bizer et al., 2011) —a set of design principles (Berners-Lee, 2006) on how to publish semantic data on the Web to be able to implement the transition from a "Web of documents to a *Web of Data*", the attention of many Semantic Web research efforts focused on the creation and publication of data on the Web, using more light-weight ontologies that exploited characteristics inherent in this kind of technology (e. g. the graph representation, the ability to create large-scale dataspaces, the possibility of dynamic typed linking, modeling taxonomic relations and reusing shared schemas), and nevertheless, relied to a lesser extent on inference. Later on, the knowledge representation facilitated by these technologies was used to model real-world knowledge in graphs, which were used by organizations under the term *Knowledge Graphs* (Hogan et al., 2021).

2.1.1 RDF Graphs

The *Resource Description Framework* (RDF) is a domain agnostic "data model for describing resources in the Semantic Web" (Antoniou et al., 2012). Developed as a W3C standard, RDF uses HTTP *Uniform Resource Identifiers* (URIs) [1] to identify resources unambiguously. These resources can be Web resources (e. g. an image in a Web site, a CSV data file) or real-world things (e. g. a person, a place) that are described on the Web. The elementary unit in RDF is a triple defined as follows:

Definition 1 ***RDF Triple****: Given $\mathcal{U}$, a finite set of HTTP URIs representing resources, $\mathcal{L}$ a finite set of literal values, and a finite set of blank nodes $\mathcal{B}$ where $U \cap \mathcal{L} = \mathcal{U} \cap \mathcal{B} = \mathcal{L} \cap \mathcal{B} = \emptyset$, an RDF triple (s, p, o) is any element of the data space $T = (\mathcal{U} \cup \mathcal{B}) \times \mathcal{U} \times (\mathcal{U} \cup \mathcal{L} \cup \mathcal{B})$.*

The (s, p, o) notation in RDF triples refers to the role that each element plays in the statement encoded in the triple, analogously to the grammatical structure common in natural languages: the first position (s) refers to the *subject*, which is the resource being described; the second position (p) is the *predicate*, which is a property that indicates the type of relation that exists between the subject and the object; and the third (o) is the *object*, that gives a value either literal or another resource. In the

An example of an RDF Triple written in pseudo-code is: `(dbr:Albert_Einstein, dbo:field, dbr:Physics)`. It describes that Albert Einstein's field of work was physics. In this case, the three elements of the triple are resources identified by URIs —two entities (in positions s and o) and the property (in position p).

Definition 2 ***RDF Graph****: a set of RDF triples constitutes an RDF Graph $G : (s, p, o)$, in which the* nodes *are resources (identified by URIs or blank nodes) and literals, and the* edges *are the typed relations established by the properties. An RDF graph is, thus, a* directed graph*, in which the relation between two elements only holds in the direction established by the triple that makes the relation explicit.*

This direction needs to be accounted when consuming the data. For instance, if an RDF graph contains the triple $(g1 : Peter, p : follows, g1 : Sean)$ the relation in the object-to-subject direction cannot be assumed, unless there is an additional triple that indicates that $(g1 : Sean, p : follows, g1 : Peter)$.

[1] URI `https://tools.ietf.org/html/rfc3986`

There are different serialization formats for the RDF data model. For instance, RDF/XML based on the XML markup language, RDFa specifically devised to embed RDF in HTML code, N3 a format in which triples are more human-readable for humans, Turtle (a subset of N3), N-Triples (a subset of Turtle), N-Quads a format including a context in each triple to indicate where the triple was specified, and JSON-LD created as an extension of JSON and considered valid JSON itself.

```
1  @prefix dbo: <http://dbpedia.org/ontology/> .
2  @prefix dbr: <http://dbpedia.org/resource/> .
3  @prefix foaf: <http://xmlns.com/foaf/0.1/> .
4  @prefix bibo: <http://purl.org/ontology/bibo/> .
5
6  # Entity 1
7  dbr:Albert_Einstein dbo:field dbr:Physics ;
8      rdf:type foaf:Person ,
9               dbo:Scientist ;
10
11     dbo:doctoralStudentOf dbr:Alfred_Kleiner ;
12     dbo:authorOf dbr:The_Evolution_of_Physics ;
13     dbo:birthDate "1879-03-1"^^xsd:date .
14
15
16 # Entity 2
17 dbr:Alfred_Kleiner dbo:almaMater dbr:University_of_Zurich ;
18                     rdf:type dbo:Scientist .
19
20 # Entity 3
21 dbr:The_Evolution_of_Physics rdf:type bibo:Book .
```

Figure 2.1: Example of an RDF graph describing Albert Einstein with Turtle notation.

Figure 2.1 provides an example of a small RDF Graph, containing 9 RDF triples, and written in the RDF Turtle notation. This RDF graph —extracted from DBpedia [2]— describes three resources/entities: the person Albert Einstein, the person Alfred Kleiner and the book “The Evolution of Physics".

In the example, Albert Einstein is declared as Person and Scientist, using the RDF construct `rdf:type`—an RDF property that classifies entities. The statements in lines 11 and 12 indicate the relationship between Albert Einstein’s entity and other entities (i.e., his PhD supervisor, and the book he co-authored), while the statement in line 13 is a case of a triple where the object is a value —the date of birth of Albert Einstein. All URIs in this example are written in a short form, using the prefixes defined in RDF. For instance, the URI `http://dbpedia.org/resource/Albert_Einstein` is abbreviated as `dbr:Albert_Einstein`, thanks to the definition of the `dbr` prefix in line 2.

Figure 2.2 provides the graph representation of the example in Figure 2.1. Entities, such as the entities referring to Albert Einstein and Alfred Kleiner are nodes in the graph, and edges indicate typed relations between the entities (e. g. `dbo:doctoralStudentOf`, `dbo:authorOf`).

RDF provides a mechanism to declared statements about statements, called RDF reification (analogous to the association class in UML representations). This instrument becomes particularly handy to express some kind of annotation information (e. g. qualifying information about the triple such as the context of its validity, or provenance information about the triple, such as the source it was extracted from and the name of the creator). In the example above, we could add reification to indicate the creator of the statement about Einstein’s field being Physics as follows:

[2] Albert Einstein’s description in DBpedia `http://dbpedia.org/page/Albert_Einstein`

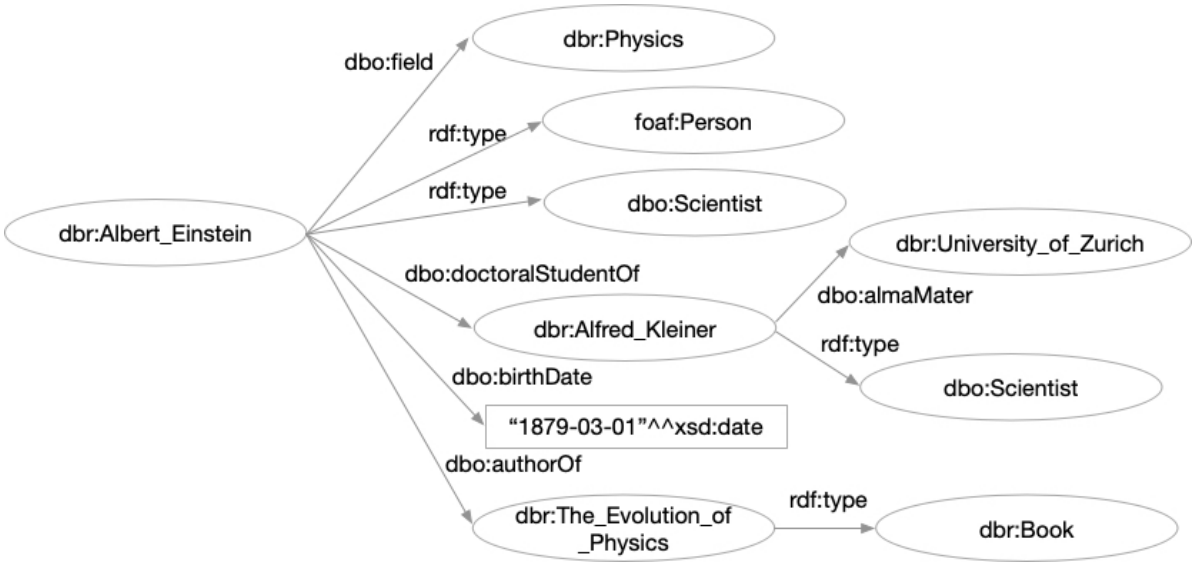

Figure 2.2: Graph representation of the previous example in Figure 2.1.

```
...
dbr:st1 rdf:type rdf:Statement ;
        rdf:subject dbr:Albert_Einstein ;
        rdf:predicate dbo:field ;
        rdf:object dbr:Physics ;
        dc:creator dbm:Curator1 .
...
```

Figure 2.3: Example of an RDF reification for one of the statements in Figure 2.1.

Over time, different RDF graphs have been developed for various applications, ranging from Web site and multimedia annotations, to enterprise data management. In the last years, a specific kind of RDF graph gained popularity: the knowledge graph —"a graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent relations between these entities" Hogan et al. (2021). While the term was originally used in Google's product called *Google Knowledge Graph*, it was adopted by scholars and organizations in industry, such as Facebook, Zalando, Microsoft and eBay, who built their own knowledge graphs to support their search, browsing and recommendation algorithms (Hogan et al., 2021).

2.1.2 Ontologies

As indicated by Guarino et al. (2009), in the field of Philosophy, concerned with characterizing reality, Ontology is related to the analysis of properties inherent in things. In Computer Science, and precisely in the context of the Semantic Web, the term has a more concrete connotation, where "ontologies are a means to formally model the structure of a system, i.e., the relevant entities and relations that emerge from its observation, and which are useful to our purposes" (Guarino et al., 2009).

Gruber et al. (1993) defined a computational ontology as the "explicit specification of a conceptualization". This widely-accepted definition, was extended by Borst (1997), who emphasized the importance of a common understanding of the conceptualization, defining an ontology as a "formal speci-

fication of a shared conceptualization". Moreover, Guarino et al. (2009) formally defined an ontology as:

Definition 3 ***Ontology****: "Let $\mathcal{C}$ be a conceptualization, and $\mathcal{L}$ a logical language with vocabulary $\mathcal{V}$ and ontological commitment $\mathcal{K}$. An ontology $\mathcal{O}_K$ for $\mathcal{C}$ with vocabulary $\mathcal{V}$ and ontological commitment $\mathcal{K}$ is a logical theory consisting of a set of formulas of $\mathcal{L}$, designed so that the set of its models approximates as well as possible the set of intended models of $\mathcal{L}$ according to $\mathcal{K}$".*

Ontologies contain the formal definition of concepts/classes and relations that are necessary to describe entities in a definite scope (i.e., for instance in the scope of computer science publications). In cases where the conceptualizations are simple (i.e., with less expressiveness), ontologies are referred to as *schemas* and *vocabularies*. In the context of Semantic Web technologies, the two main ontology specification languages are RDFS (RDF Schema) (Guha and Brickley, 2014) and OWL 2 Web Ontology Language (Hitzler et al., 2009).

Both languages allow to define classes, properties and hierarchies of both, but each language provides different constructs to do so. In RDFS, for instance, classes are defined as `rdfs:Class`, properties as `rdfs:Property` and subsumption is specified with `rdfs:subClassOf` and `rdfs:subPropertyOf`. An example of a class is `Dish` and `VegetarianDish` is an example of a subclass. In OWL, classes are defined as `owl:Class` and for properties, there is a distinction between properties used to connect entities (`owl:objectProperty`) and properties used to provide a literal value (`owl:datatypeProperty`). Subsumption is defined via RDFS subclass/subproperty constructs. OWL, in contrast to RDFS, allows to define class disjointness, class and property equivalence, as well as inverse, symmetric and transitive properties. OWL additionally contains annotation properties that can be used for documentation purposes.

Based on the scope's specificity and the purpose served by the ontology, different types of ontologies can be distinguished: *upper ontologies* (also known as foundational or top-level ontologies) specify a general cross-domain conceptualization, while *domain-specific ontologies* provide the definition of concepts and relations that relate to a more narrow part of the world. DOLCE (Gangemi et al., 2002) is an example of an upper ontology. This ontology defines abstract concepts to distinguish between so-called *endurants* (i.e., "things that are in time") and *perdurants* (i.e., "things that happen in time") (Gangemi et al., 2002). It also defines concepts and relations to model temporal and physical qualities, social and mental objects, as well as events and processes. Such an upper-ontology is often used to integrate multiple heterogeneous ontologies under the umbrella of a more general one (i.e., the upper-ontology). A domain ontology, instead, defines specific classes and relations. For instance, the Music Ontology [3] defines classes to model instruments, music festivals, records and artists.

Looking again at the example in Figure 2.1, the RDF triples use elements defined in ontologies: FOAF (Brickley and Miller, 2004), the DBpedia ontology [4] and SIOC (Berrueta et al., 2007). The DBpedia ontology defines the class `dbo:Scientist`, used to type the entity (or RDF resource) referring to Albert Einstein, and the relation `dbo:authorOf`, used to connect the entity of Einstein to the entity referring to a book he authored.

In the last years, several ontology engineering methodologies have been defined (Tudorache, 2020), including the simple but effective method by Noy and McGuinness (2001), the NeON methodology (Suárez-Figueroa et al., 2012) and the UPON Lite methodology (Nicola and Missikoff, 2016). One initial task in the ontology engineering process, common to the majority of existing methodologies is

[3] Music Ontology Specification `http://musicontology.com/specification/`

[4] DBpedia ontology `https://wiki.dbpedia.org/services-resources/ontology`

the definition of the scope of the ontology (i.e., defining the future use of the ontology and its domain). An effective way to define the scope is by answering so-called competency questions —questions that help identify the information that the ontology should contain, thus, clarify the needs for specific classes and properties. Another best practice acknowledged by many methodologies and supported by ontology engineering frameworks is to reuse existing ontologies. RDF technologies facilitate this task, as it is possible to simply include the URIs of existing classes and properties of third-party ontologies, and link to them. Reusing widely-adopted ontologies helps to increase the interoperability of the data using the ontology/es. Besides reusing concrete ontologies, it is also possible to reuse ontology design patterns (ODPs) (Gangemi, 2005) —which propose best modeling practices similarly to the design patterns in software engineering (Gamma et al., 1994).

While some ontologies can be technically designed and implemented by a single ontology engineer, it is desirable to involve multiples knowledge engineers, as well as domain experts in the process. The DILIGENT methodology (Pinto et al., 2009) is an example of such a distributed process, that acknowledges that ontologies may evolve over time and facilitates an argumentation framework to discuss modeling decisions.

2.1.3 SPARQL

SPARQL is the W3C standard language for querying RDF graphs (Prud'hommeaux et al., 2013). SPARQL is also a protocol that defines the way SPARQL queries are conveyed and results are sent over HTTP, to/from a SPARQL endpoint or service.

Having some commonalities with SQL in terms of basic formulations for selection and projection, SPARQL queries define conditions to consult the data. In order to do so, the SPARQL syntax uses the primary RDF constructs (URIs and triples). There are different types of SPARQL queries:

1. ***SELECT* queries** specify graph patterns containing variables and fixed values (URIs/literals), in order to establish the conditions that the results of the query should match. The result of a `SELECT` query is the set of variables that are bound to values based on the graph pattern matching. Figure 2.4 shows an example of a simple `SELECT` SPARQL query that obtains from DBpedia the URI of the field of work of scientists. The graph pattern used to find relevant resources sets two requirements in two triple patterns: that the resource has been explicitly typed as `dbo:Scientist` and a field of work has been declared.

```
PREFIX dbo: <http://dbpedia.org/ontology/>.
PREFIX dbr: <http://dbpedia.org/resource/>.

SELECT ?scientist ?field
WHERE{
    ?scientist rdf:type dbo:Scientist .
    ?scientist dbo:field ?field .
}
```

Figure 2.4: Example SPARQL SELECT query.

2. ***ASK* queries** specify graph patterns, but in contrast to `SELECT` queries, give as a result a boolean value (true/false) indicating whether the graph pattern is matched in the dataset or not. Figure 2.5 shows an example of an `ASK` query that checks whether DBpedia has contains the date of birth of Albert Einstein.

```
PREFIX dbo: <http://dbpedia.org/ontology/>.
PREFIX dbr: <http://dbpedia.org/resource/>.


ASK { dbr:Albert_Einstein dbo:birthDate ?date }
```

Figure 2.5: Example SPARQL ASK query.

3. ***CONSTRUCT* queries** aim at building an RDF (sub-)graph from the dataset that is queried. Hence, these queries specify graph patterns and return RDF triples that match the graph patterns. Figure 2.6 shows an example of a `CONSTRUCT` query, that creates a graph with the publications of scientists working in the field of Physics.

4. ***DESCRIBE* queries** are used to obtain the RDF description of resources, which can be set by the URI or via a graph pattern. Figure 2.7 shows an example to obtain the RDF description of Albert Einstein from DBpedia. The resultset may contain the complete set or a subset of the resource's triples in the dataset depending on the implementation of the infrastructure hosting

```
1 PREFIX dbo: <http://dbpedia.org/ontology/> .
2 PREFIX dbr: <http://dbpedia.org/resource/> .
3
4 CONSTRUCT { ?publication dbo:author ?scientist }
5 WHERE {
6     ?scientist rdf:type dbo:Scientist .
7     ?scientist dbo:field dbr:Physics .
8 }
```

Figure 2.6: Example SPARQL CONSTRUCT query.

and querying the data, as some implementations allow to set an upper-bound in the number of triples to return.

```
1
2 DESCRIBE <http://dbpedia.org/resource/Albert_Einstein>
```

Figure 2.7: Example SPARQL DESCRIBE query.

All triple patterns in a SPARQL query's graph pattern need to be matched, unless the `OPTIONAL` keyword is used to signal that the matching of the triple pattern(s) within the scope of the `OPTIONAL` can be relaxed. For example, in the former `SELECT` query, we could add a triple pattern to obtain the URI of the people the scientists influenced. However, this information might not be specified for some scientists, and we would not like the scientists to be removed from the resultset for that reason, because we would still interested in their birth date. Hence, we would add the triple pattern with optional matching as " `OPTIONAL { ?scientist dbo:influenced ?person } .` ".

In order to specify disjunctive graph patterns, SPARQL has the `UNION` keyword. For instance, if we would like to select people classified as the class `Person` in the DBpedia ontology or the class `Person` in the FOAF ontology, we would add the following to the graph pattern: `?person rdf:type dbo:Person UNION ?person rdf:type foaf:Person`. Additionally, one can state negation in graph patterns to select resources that do not have a specific triple pattern (with `FILTER NOT EXISTS`), or to remove matches based on a triple pattern (with `MINUS`). For further filtering based on value-based constraints (e. g. regular expressions), SPARQL queries can include `FILTER` statements. Furthermore, SPARQL provides a very powerful mechanism dubbed *property paths* to specify paths with several edges. This feature is very useful for queries that entail looking up taxonomic relations such as subsumption (e. g. querying for a specific type of resources or any resource classified as a subclass of such type).

Similarly to SQL, SPARQL provides solution modifiers to order results (`ORDER BY`), restrict the number of results (`LIMIT`), start after a specific number of solutions (`OFFSET`) or obtain unique values (`DISTINCT`). Aggregates can be created with GROUP BY and refined via the `HAVING` construct. Moreover, aggregate functions can be used to compute simple counts (`COUNT`), the sum (`SUM`), the minimum/maximum values ((`MIN` and `MAX`)) and a concatenation (`GROUP_CONCAT`) with the `GROUP BY` and `HAVING` modifiers.

Figure 2.8 shows an example of a query using several of the options listed above. This query selects the URI of people, the URI of their image (if it is present) and the count of publications that the person authored. The resultset is limited to 50 and order in descending order in terms of the count of publications. In order to query for people, we look for the resources classified as `dbo:Person` or any more specific class defined as a direct or indirect subclass of `dbo:Person`.

```
PREFIX dbo: <http://dbpedia.org/ontology/>.
PREFIX dbp: <http://dbpedia.org/property/>.
PREFIX dbr: <http://dbpedia.org/resource/>.

SELECT ?person ?image (count(distinct ?publication) as ?count)
WHERE{
  ?person rdf:type/rdfs:subClassOf* dbo:Person .
  ?publication dbo:author ?person .
  OPTIONAL{?person dbp:image ?image .}
}
GROUP BY ?person ?image
ORDER BY DESC (?count)
LIMIT 50
```

Figure 2.8: Example SPARQL query.

2.1.4 Linked Data

The *Linked Data* principles (Berners-Lee, 2006) aim to establish a list of requirements for data to be leveraged into a semantic dataspace. They suggest that:

1. Things/resources should be identified with HTTP URIs on the Web.
2. The description of things/resources should be made available by means of (Semantic Web) standards (i. e. RDF and related languages for structured data annotation and its consumption).
3. Resources should be *dereferenceable*: an HTTP GET request of the URI identifying a thing should return its RDF-based description
4. Resources should be linked to other resources: the description of a resource should contain the relation to other resources, so that more things can be discovered.

The W3C *Linking Open Data* initiative promoted the application of these principles to open data (i. e. data that can be used, altered and re-shared for free). As a result, the so-called "LOD cloud" was created —a set of 1.4K interconnected datasets published conforming to the Linked Data principles. The principles were also slightly reformulated as a 5-star scheme[5] for Linked Open Data, that suggests data to be: (i) published on the Web under an open license, (ii) machine-readable or structured data, (iii) in non-proprietary format, (iv) implemented with RDF standards, and (v) linked RDF.

Among the publishers of these LOD datasets are media outlets such as the New York Times and the BBC, large academic efforts involving several institutions and research projects, as well as GLAM (Galleries, Libraries, Archives and Museums) and startups. The domain of these datasets varies from geography and publications, to social networking, linguistics, life sciences, governmental and user-generated data.

While the Linked Data principles were defined around data, they equally apply to ontologies/vocabularies. The LOV Web portal (Vandenbussche et al., 2017a) aggregates and describes a set of over 700 linked open vocabularies that have been specified for a wide variety of domains [6].

Despite the increase in the volume of Linked Data, not all of these datasets have full-time availability and some have stopped their service —as it can be seen through the SPARQLes monitoring tool

[5] 5-Star LOD `https://5stardata.info/en/`

[6] Linked Open Vocabularies `https://lov.linkeddata.es/dataset/lov/`

(Vandenbussche et al., 2017b). Many of these LOD datasets have supplied data to end-user applications, and others have been extensively used for the scientific evaluation of semantic data management methods. One of the most prominent datasets resulting from this effort is DBpedia, which contains cross-domain human knowledge automatically extracted from Wikipedia Infoboxes. DBpedia became the central dataset in the cloud, also named “interlinking hub", as it has a large volume of incoming and outgoing links from/to resources of other datasets.

In order to facilitate the programmatic discovery and analysis of datasets, LOD dataset descriptions are also published in a machine-readable manner. For this purpose, the data portal hosting the LOD datasets uses the DCAT vocabulary[7]. This vocabulary defines the classes and properties needed to indicate a dataset’s creator, description, license and access details, among other things. Google’s data search engine, currently indexing millions of datasets, relies on DCAT and Schema.org data annotations of data to facilitate people find datasets (Brickley et al., 2019). VoID[8] is another vocabulary for describing datasets. It contains classes and properties to define information that appears specifically in Linked datasets (and not in other types of datasets), such as linksets (i. e. groups of links from a source dataset to a target dataset).

[7] DCAT `https://www.w3.org/TR/vocab-dcat-2/`
[8] VoID `https://www.w3.org/TR/void/`

2.2 Semantic Web Data Integration

The key task in Semantic Web data integration, and in dataspaces in general, is to connect or *link* the items of individual data sources. While in the Semantic Web it is possible to link sources using links associated to any well-defined meaning, traditionally this linkage has been implemented with regard to *similarity*. Similarity can be related to equivalence or identity, depending on the nature of the items to be linked —concepts/properties of a schema, or entities of a dataset respectively.

Over the years, the scientific literature in data management has coined different terms for this task, namely *ontology/schema matching* and *ontology alignment* when it refers to ontologies, and *data matching*, *record linkage*, and *entity resolution* when it relates to data entities (Christen, 2012).

For simplicity, when I provide definitions and describe linking processes in the remainder of this section, I refer to “links" and “link discovery" in relation to RDF graphs, abstracting from the nuances of item types (classes/properties/entities). I also refer to RDF datasets, which are graphs that follow the Linked Data recommendations (Hogan et al., 2021). Hence, the reader may interpret the term link as an ontology mapping or a link between two entities. In parts where the differentiation is relevant (e. g. in Subsection 2.2.3) I separately mention ontology mapping and data link discovery.

2.2.1 Links

In order to provide a formal definition of a link, let us extend the aforementioned definition of RDF triple to RDF quadruple, to introduce the notion of context that is necessary to define boundaries in Linked Data.

Definition 4 ***RDF Quadruple****: Given $\mathcal{U}$, a finite set of HTTP URIs, representing resources, $\mathcal{L}$ a finite set of literal values, and a finite set of blank nodes $\mathcal{B}$ where $U \cap \mathcal{L} = \mathcal{U} \cap \mathcal{B} = \mathcal{L} \cap \mathcal{B} = \emptyset$, a quadruple (s, p, o, c) is any element of the data space $Q = (\mathcal{U} \cup \mathcal{B}) \times \mathcal{U} \times (\mathcal{U} \cup \mathcal{L} \cup \mathcal{B}) \times \mathcal{U}$. s, p, o is a triple statement describing s, while c is the context (denoted by a URI) in which the triple is defined.*

Definition 5 ***RDF Dataset with Context****: An RDF dataset D_c is a set of quadruples grouped by some context c $D_c \subseteq \{(s, p, o, c) \in Q\}$, where Q is the set of all quadruples.*

Definition 6 ***Home****: Given C the set of all contexts, and an entity (either a blank node or URI), $home : \mathcal{B} \cup \mathcal{U} \mapsto C$ is the function that maps the entity to the context c where the entity is defined. Note that when x is a vocabulary term (e. g. a class or a property), the c returned by $home(x)$ is the identifier of the vocabulary where the term x was defined.*

Definition 7 ***Link****: A link of D_c is a quadruple $(s, p, o, c) \in D_c$ such that $s \in \mathcal{U}, o \in \mathcal{U}, home(s) = c, home(o) \neq c$.*

Therefore, a link is the triple that connects items of two different contexts. The set of links between two concrete sets of items grouped by context is called *alignment* or set of mappings in the case of ontologies, and *linkset* in the case of data graphs. Again, for simplicity let us refer to a common term:

Definition 8 ***Interlinking****: The interlinking I_c of an RDF dataset D_c is the set of all links going out from D_c to any RDF dataset: $I_c = \{(s, p, o, c) \in D_c |\ home(s) = c, home(o) \neq c\}$.*

2.2.2 Link Discovery: Task Definition

The task of link discovery is the task of finding the links (or correspondences) between the items of individual graphs (Euzenat et al., 2007). Hence, given two RDF graphs/datasets to be linked via a concrete type of predicate, the task of link discovery can be specified as follows:

Input:

- *Source RDF Graph/Dataset*: RDF graph/dataset where the interlinking should originate from. Resources of this graph appear in the *subject* position of links.
- *Target RDF Graph/Dataset*: RDF graph/dataset where the interlinking should end at. Resources of this graph appear in the *object* position of links.

Procedure: Despite differences in the way specific solutions solve the task of link discovery, a general procedure for link discovery can be identified according to the steps listed by Christen (2012):

- *Pre-processing*: the first step is to ensure that the input has descriptions of resources that are of high quality (i. e. without misspellings and unwanted words, correct values) and easy to compare (i. e. all values have the same lower-/upper-case conventions and are specified in a standard or well-defined manner). Therefore, this step involves data cleaning and data transformation actions.
- *Blocking (aka indexing)*: let R_S be the set of resources in the source RDF graph/dataset and R_T the resources in the target RDF graph/dataset, the space of possible links to be in the interlinking is the Cartesian Product (CP) $R_s \times R_T$. However, the majority of these matches would be incorrect, and calculating the CP of resources in large RDF graphs/datasets does not scale. The goal of this step is to reduce the number of matches to be analyzed according to a blocking criteria defined in terms of the knowledge modeled in the RDF graph/dataset (e. g. compare only publications that share the field of research). The result of this step is a first set of *candidate links*.
- *Item pair comparison*: each pair of resources involved in candidate links needs to be compared. This comparison can be at three different levels: *syntactic* (e. g. comparing name strings), *structural* (e. g. comparing the graph structure of entities) or *semantic* (e. g. comparing the logical definition of entities).
- *Classification*: in this step the link discovery algorithm may decide for each *candidate link*, if it is a *clear link*, a *clear no link* or an *uncertain case*.

Output:

- Interlinking between the source and the target RDF graphs/datasets.

Direction Note that while, as defined, the interlinking is sensitive to direction (from source to target), the link discovery process can result in bidirectional links (i. e. using symmetric properties as link predicate or generating inverse links in addition) in order to facilitate the consumption of the integrated data in both directions.

Link Predicates There is a wide variety of defined properties that can be used as link predicates. Figure 2.1 gives a summary of the type of relationships that links can express.

First, there are relationships that apply to ontology elements:

- *Equivalence* implies that two classes or two properties serve the same function; instances of one of the two classes become instances of the other class.
- *Subsumption* indicates that one class/property is more specific or more general than another class/property.
- *Disjointness* expresses that there cannot be an instance that is of type of both classes at the same time.

Second, there are relationships that are meant to relate entities in datasets:

- *Identity* indicates that two entities denote the exact same thing in the real world.
- *Similarity* signals that two entities show close similarity while being different things in the real world.
- *Domain-Relationship* any other kind of relationship defined in a domain ontology can also be employed to connect datasets.

Last, both ontology items and data entities may be linked in a more vague (but still useful) manner with more generic predicates that encourage the data consumer to look up other connected resources.

Schema Link Predicates	
Equivalence	`owl:equivalentClass`, `owl:equivalentProperty`
Subsumption	`rdfs:subClassOf`, `rdfs:subPropertyOf`
	`skos:broadMatch`, `skos:narrowMatch`
Disjointness	`owl:disjointWith`
Data Link Predicates	
Identity	`owl:sameAs`, `skos:exactMatch`
Similarity	`skos:closeMatch`, `skos:relatedMatch`
Domain Relationship	`wgs84:location`, `foaf:knows`
Link Predicates for Both	
General	`rdfs:seeAlso`

Table 2.1: List of Predicates Used in Schema and Data Links.

Complexity The types of relationships listed above establish links between so-called *simple expressions* (i. e. single resources). A relatively recent body of work in ontology alignment has looked into methods to discover more complex alignments that connect resources through *complex expressions*, which are “composed of at least one entity on which a constructor or a transformation function is applied" (Thiéblin et al., 2019). Moreover, Beek et al. (2016) proposed that the validity of identity links can be context-dependent.

2.2.3 Link Discovery Algorithms in the Semantic Web

In the last two decades, a large number of algorithms, frameworks and tools were implemented for link discovery in RDF graphs. The literature distinguishes methods for ontology alignment (Pavel and Euzenat, 2013) from methods for (entity) link discovery (Nentwig et al., 2017). While the latter operate with data of a larger scale and smaller logical complexity, and therefore, use slightly different techniques, the methods of both scenarios evolved over the years increasing the level of automation. Initial frameworks for link discovery required users (i.e., data publishers) to implement a set of rules that indicated the type of entities to be linked, the attributes to be compared in the comparison phase, a similarity threshold, and the predicate to use in the link triples. Once these rules were specified in XML, the framework would parse and apply them to generate the link triples as output.

As indicated in Section 2.2.2, there are numerous predicates that can be used for specifying links between classes, links between properties and links between entities/instances. Nonetheless, most of the related work focuses on identity and equivalence, as these relationships can be defined in a generic way, and the resulting links facilitate *join*-like queries. Existing link discovery algorithms implement the ontology/data comparison step focusing on different characteristics of the data: namely, *syntactic*, *structural* and *semantic* characteristics (Pavel and Euzenat, 2013; Nentwig et al., 2017). A syntactic analysis of RDF graphs entails the comparison of literal values that can either be strings referring to the names of entities or classes, or any other literal that appears in the triples of the entity's/class' RDF description. String similarity measures, such as the Levenshtein measure or the Jaccard measure are often used to perform syntactic analysis (Christen, 2012). In contrast, structural analysis focuses on the structure of the RDF (sub)graph and the relations of the entity/class at hand with other entities/classes. A structural analysis takes into account subclasses, superclasses, siblings and connected entities. Finally, a semantic analysis of the RDF graphs takes into account logical restrictions specified in RDF. For instance, such an analysis acknowledges the definition of functional properties and disjoint classes.

With the increasing improvement of machine learning techniques, entity link discovery algorithms, such as Silk and LIMES (Nentwig et al., 2017), implemented an active learning (Settles, 2012b) approach to solve the linking task. Active learning (AL) is a supervised learning technique in which the algorithm selects data points (i.e., candidate links in this case) to be labelled by an oracle (i.e., a human expert labeller who is able to provide a ground truth). Depending on the *query strategy* defined by the AL algorithm, the next candidate link to be labelled can be selected differently. Some of the most common query strategies include uncertainty sampling (i.e., selecting the candidate links for which the algorithm is less certain), and query by disagreement (i.e., selecting the candidate links for which an ensemble of possible linking rules disagree the most).

While there are solutions that aim to connect more than two RDF graphs at once (e.g., the holistic ontology alignment algorithm by Bock et al. (2012)), given the complexity of the problem, the majority of frameworks connect two RDF graphs.

Evaluating Link Discovery Algorithms Link discovery algorithms tend to be evaluated with Information Retrieval measures: using *precision*, *recall* and *F1* (Manning et al., 2008). The Ontology Alignment Evaluation Initiative [9] uses these measures in its benchmarking campaigns for assessing state-of-the-art ontology alignment and entity link discovery (aka instance macthing) algorithms.

Given a set of links identified by an algorithm, and a ground truth (i.e., reference data that describes the knowledge of the real-world and hence indicates if links are true or not), these measures help identify if all the links in the ground truth were found by the algorithm (recall), and how many links

[9] OAEI `http://oaei.ontologymatching.org/2020/`

		Ground Truth: Link	Ground Truth: No Link	Total
Interlinking to be Evaluated	Link	TP	FP	$TP + FP$
	No Link	FN	TN	$FN + TN$
	Total	$TP + FN$	$FP + TN$	N

Table 2.2: Confusion Matrix.

from all those found by the algorithm are correct (precision). F1 is a combination of precision and recall.

Table 2.2 shows the definition of a confusion matrix, which is used to quantify the errors (and correct matches) made by link discovery algorithms. The matrix counts:

- the number of candidate links that were identified as links by the algorithm and should be links according to the ground truth (True Positives, or in short TP).
- the number of candidate links that were identified as links by the algorithm and should not be links according to the ground truth (False Positives, or in short FP).
- the number of candidate links that were identified as no links by the algorithm and should be links according to the ground truth (False Negatives, or in short FN).
- the number of candidate links that were identified as no links by the algorithm and should not be links according to the ground truth (True Negatives, or in short TN).

Based on these values, precision, recall and F1 are then defined as follows (Manning et al., 2008):

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

2.2.4 Making Use of Links While Consuming Linked Data

Linking items of individual and distributed RDF graphs facilitates their joint consumption, as it creates the structure that enables a direct navigation between them. However, in order to access multiple RDF graphs and obtain cross-RDF graph results in a seamless manner, an integrative look-up mechanism is necessary. Two main strategies can be distinguished to implement a solution to this problem: *federated query* and *automated link traversal*. In federated query (Görlitz and Staab, 2011; Aranda et al., 2014), an initial query is split into subqueries that a federation component distributes to the individual sources hosting the RDF graphs, collects the individual results and delivers them in an aggregated way. This approach relies on SPARQL endpoints serving the individual graphs. Instead, in query solutions that rely on automated link traversal (Hartig, 2013; Umbrich et al., 2015), the data is mainly consumed via a Linked Data interface. The query is executed over a dataset that iteratively grows by adding new data discovered by traversing data links.

SPARQL 1.1 includes an extension for query federation (Prud'hommeaux et al., 2013). As showcased in the example in Figure 2.9, SPARQL uses the keyword *SERVICE* to indicate the URI of the (external) SPARQL endpoint to be accessed —in this case DBpedia's and Wikidata's SPARQL endpoints. This query obtains the URI of entities typed as "scientist" in DBpedia and the URI of the field they work at, and additionally looks up the native language of the scientists in Wikidata, using an identity link that matches the URI of each scientist in DBpedia to their URI in Wikidata.

```
1  PREFIX dbo: <http://dbpedia.org/ontology/> .
2  PREFIX dbp: <http://dbpedia.org/property/> .
3  PREFIX dbr: <http://dbpedia.org/resource/> .
4
5  PREFIX wd: <http://www.wikidata.org/entity/> .
6  PREFIX wdt: <http://www.wikidata.org/prop/direct/> .
7
8  SELECT DISTINCT ?scientistDB ?field ?nativeLang
9  WHERE
10   {
11
12   SERVICE <http://dbpedia.org/sparql>
13       { ?scientistDB rdf:type dbo:Scientist ;
14                   dbo:field ?field ;
15                    owl:sameAs ?scientistWD .
16
17          }
18     SERVICE
           <https://query.wikidata.org/bigdata/namespace/wdq/sparql>
19       { ?scientistWD wdt:P103 ?nativeLang
20       }
21
22   }
```

Figure 2.9: Example SPARQL federated query.

2.3 Semantic Data Quality Assurance

Data quality assurance is the process in which data is assessed, and possibly modified for improvement, in terms of a (set of) concrete dimension(s), using specific measures and in the context of (a) reference value(s) (Batini et al., 2016). This data quality process is crucial to ensure the intended performance of information systems. There are numerous examples that illustrate the negative consequences that low-quality data can lead to. NASA's Mars Climate Orbiter [10], for instance, was destroyed due to the incorrect specification of a measurement, which was supposed to be in SI units (i. e. International System of Units) indicating the number of Newton seconds, but was instead specified in "pound-force seconds" (a non-SI unit). After the incident, the team responsible for the mission acknowledged that they should have followed the procedures for checking the correctness of their data more strictly (Oberg, 1999). In a different scenario, when routing algorithms use outdated geo-locations of points of interests, they send users to a destination that is different from the one users look for. Furthermore, with the rise of Data Science methods, there are millions of decisions that are made within industrial and governmental contexts in the interest of economic growth, resource planning and society's well-being,

[10] Wikipedia Article for NASA's Mars Climate Orbiter https://en.wikipedia.org/wiki/Mars_Climate_Orbiter

which rely on high quality data. Hence, ensuring that data has adequate properties is very important for a wide variety of tasks and circumstances.

After Linked Data gained momentum in the Semantic Web field (Heath and Bizer, 2011), semantic data quality assurance became one of the imperative challenges to be addressed (Bernstein et al., 2016). As a result, the literature contains numerous works that refer to semantic data quality definitions, methodologies, as well as the implementation of frameworks and tools to (i) assess, and to a lesser extent also (ii) to improve and monitor the evolution of data quality in RDF data:

Semantic Data Quality Dimensions Zaveri et al. (2016) surveyed existing methods for Linked Data quality assessment adopting the well-established multi-dimensional data quality framework defined by Wang and Strong (1996). For each dimension, the authors reviewed the literature on related measures and systems tailored to RDF data published as Linked Data. Moreover, the authors introduced four relevant dimensions that were not present in the work by (Wang and Strong, 1996): licensing, interlinking, performance and versatility.

Both works (Wang and Strong, 1996; Zaveri et al., 2016) categorize data quality dimensions into four major categories: *intrinsic dimensions*, *contextual dimensions*, *representational dimensions* and *accessibility dimensions*. Intrinsic dimensions are inherent in the data, independent of the context in which the data is used, in contrast to contextual dimensions, which depend on the use case or scenario in which the data is used. Representational dimensions refer to properties related to the way the data is specified and interpreted, whereas accessibility dimensions refer to the way the data can (or cannot) be accessed.

In the following, a short introduction to these data quality dimensions is presented in the context of Linked Data.

Intrinsic dimensions :

- **Syntactic Validity** refers to the extent "to which an RDF document conforms to the specification of the serialization format" (Zaveri et al., 2016). *Example:* An unclosed XML tag in an RDF/XML document makes the document syntactically invalid.
- **Semantic Accuracy** is "the degree to which data values correctly represent the real world facts" Zaveri et al. (2016). *Example:* When an RDF statement indicates that Barcelona is located in Germany there is semantic inaccuracy, because that city's country in the real world is Spain.
- **Consistency** "means that a knowledge base is free of (logical/formal) contradictions with respect to particular knowledge representation and inference mechanisms" (Zaveri et al., 2016). *Example:* An ontology may indicate that the classes `Student` and `Employee` are disjoint, while a dataset instantiating the ontology contains the description of an RDF resource that is declared of type Student and of type Employee. The data violates a logical constraint and is, hence, inconsistent.
- **Completeness** "refers to the degree to which all required information is present in a particular dataset. In terms of Linked Data, completeness comprises of the following aspects: (i) Schema completeness, the degree to which the classes and properties of an ontology are represented, thus can be called "ontology completeness", (ii) Property completeness, measure of the missing values for a specific property, (iii) Population completeness is the percentage of all real-world objects of a particular type that are represented in the datasets and (iv) Interlinking completeness, which has to be considered especially in Linked Data,

refers to the degree to which instances in the dataset are interlinked" (Zaveri et al., 2016). *Example:* As an example of description completeness, one can imagine that a scholarly dataset will contain complete instance descriptions if for each entry, the description includes values for all mandatory BIBTEX attributes. A knowledge graph is complete in terms of city population, when it describes every city officially recognized by official governmental authorities.

- **Conciseness** " refers to the minimization of redundancy of entities at the schema and the data level. Conciseness is classified into (i) intensional conciseness (schema level) which refers to the case when the data does not contain redundant schema elements (properties and classes) and (ii) extensional conciseness (data level) which refers to the case when the data does not contain redundant objects (instances)" (Zaveri et al., 2016). *Example:* An ontology that contains two equivalent and redundant properties for labeling books (e. g. `name` and `title`) is less concise than the same ontology having only one single property for this purpose.

Contextual Dimensions :

- **Relevancy** "refers to the provision of information which is in accordance with the task at hand and important to the users' query" (Zaveri et al., 2016). *Example:* In a scholar dataset to be used by an application like DBLP [11], having data that describes private activities of the authors is irrelevant.
- **Understandability** "refers to the ease with which data can be comprehended without ambiguity and be used by a human information consumer" (Zaveri et al., 2016). *Example:* Classes, properties and instances in RDF graphs have machine-readable names or identifiers, which can be a numerical value (e. g. `Q72` is the identifier for the instance representing the city of Zurich in Wikidata). Human-readable labels, such as the German name of the city (Zürich), help recognize and understand the entity the RDF description refers to.
- **Timeliness** "measures how up-to-date data is relative to a specific task" (Zaveri et al., 2016). *Example:* Some data, such as the name of the president of a country, tends to change over time and is expected to be up-to-date when consumed.
- **Trustworthiness** "is defined as the degree to which the information is accepted to be correct, true, real and credible" (Zaveri et al., 2016). Objectivity, as well as the reputation of the source publishing the data are considered key factors for trustworthiness. *Example:* The trustworthiness of data produced by governmental statistical offices is usually higher than crowdsourced data.

Representational Dimensions :

- **Interpretability** "refers to technical aspects of the data, that is, whether information is represented using an appropriate notation and whether the machine is able to process the data" (Zaveri et al., 2016). *Example:* There are numerous data format conventions that make temporal and local data more interpretable. For instance, a reference to a country will be more interpretable when it is written as an ISO country code (ISO, 2021), than when it is provided in a local language.

[11] DBLP `https://dblp.uni-trier.de/`

- **Interoperability** "is the degree to which the format and structure of the information conforms to previously returned information as well as data from other sources" (Zaveri et al., 2016). *Example:* A dataset whose entities are described using a vocabulary used by many applications (e. g. Dublin Core) is more interoperable than the same dataset using instead an own vocabulary defined from scratch.
- **Versatility** "refers to the availability of the data in different representations and in an internationalized way" (Zaveri et al., 2016). *Example:* A dataset with content exclusively in German and shared as an SQL file is less versatile than a dataset with text translated to all European language and shared as JSON, CSV, SQL and RDF/XML files.
- **Representational Conciseness** "refers to the representation of the data, which is compact and well formatted on the one hand and clear and complete on the other hand" (Zaveri et al., 2016). *Example:* An RDF reified statement has a more complex representation than a simple statement.

Accessibility Dimensions :

- **Licensing** "is defined as the granting of permission for a consumer to re-use a dataset under defined conditions" (Zaveri et al., 2016). *Example:* Several licenses exist for open datasets, including those from Creative Commons and Open Data Commons. The CC0 1.0 license [12] allows consumers to use the data as public domain.
- **Availability** "is the extent to which data (or some portion of it) is present, obtainable and ready for use" (Zaveri et al., 2016). *Example:* Despite being published, data may be temporarily unavailable when the server hosting the data experiences technical problems.
- **Interlinking** "refers to the degree to which entities that represent the same concept are linked to each other, be it within or between two or more data sources" (Zaveri et al., 2016). *Example:* As mentioned above, linking entities is one of the core requirements for data to be considered Linked Data. A prominent example of this dimension is the identity linkage existing between DBpedia and Wikidata entities.
- **Security** "is the extent to which data is protected against alteration and misuse" (Zaveri et al., 2016). *Example:* Like any software system, Linked Data applications and servers need to implement security mechanisms to preserve the data as it was created and avoid malicious third-party actions. As mentioned by Zaveri et al. (2016), digital signatures are an example mechanism to be used with this regard.
- **Performance** "refers to the efficiency of a system that binds to a large dataset, that is, the more performant a data source is the more efficiently a system can process data" (Zaveri et al., 2016). *Example:* Depending on the concrete characteristics of the applications using the data, performance might need to be measured in different way. While query response time is definitely important in the vast majority of systems, popular endpoints receiving the attention of a high volume of users/applications, like Wikidata, also need to provide (i. e. the number of (query) requests that can be successfully processed per unit of time).

As indicated in some of the aforementioned dimensions, Linked Data quality relates to both the schema and the data itself. The field of ontology engineering, though, dedicated parallel extensive efforts to define ontology evaluation measures and methodologies (Vrandecic, 2010). Some of these

[12] CC0 1.0 license {https://creativecommons.org/publicdomain/zero/1.0/}

works adopted quality dimensions from data quality frameworks (e.;g. consistency, completeness, conciseness), and extended them with ontology-specific dimensions, such as expandability, adaptability and mappability (i. e. whether an ontology "can be mapped to upper level or other ontologies") (Gómez-Pérez, 2004; Obrst et al., 2007). OOPS!, the ontology pitfall scanner, provides a tool to identify ontology quality issues classified into "structural, functional, and usability pitfalls" (Poveda-Villalón et al., 2014).

These Linked Data quality dimensions can be addressed using different measures that may focus on different aspects of the data, or may implement different computations. For example, completeness can be measured as a relative or absolute completeness. For an exhaustive description of existing measures and assessment methods, the reader is encouraged to refer to the original survey (Zaveri et al., 2016).

Linked Data Quality Best Practices and Tools The W3C published a recommendation on publishing human- and machine-readable data on the Web (Greiner et al., 2017). While the Linked Data principles are at the core of these best practices, the W3C's recommendation specifically suggest to include RDF-based metadata describing information about licensing, provenance, access and data quality of the dataset. In order to describe the quality of the data, the Data Quality Vocabulary (DQV) (Albertoni et al., 2016) is recommended to be used —a schema containing classes and properties to specify quality values, the measures employed to compute such values, and the dimensions these measures address. Additionally, the ontology allows to specify any quality certificate obtained for the dataset. The DQV vocabulary is also the basis of the Luzzu framework by Debattista et al. (2016). Luzzu implements a set of well-known measures for Linked Data dimensions and facilitates the addition of further implementations. It provides a GUI that allows users to explore the quality values at a certain point in time, its evolution and a ranking of datasets based on configurable weighted quality measurements.

Beek et al. (2014) developed a framework that collects, cleans and re-publishes data conforming to the Linked Data principles. The framework, which is able to process billions of triples, crawls RDF data from the Web, checks for syntactic errors (e. g. incorrect encoding, illegal characters, tags that do not match), implements statement de-duplication, serializes the data in a uniform format, programmatically generates VoiD dataset descriptions and re-publishes the data.

More recently, SHACL —a language for shape expressions— was defined to validate the extent to which an RDF graph satisfies a set of constraints pertaining the structure and the content of the data (Knublauch and Kontokostas, 2017). Shape expressions help identify, for example, missing, incorrect and outdated data, as well as misstyped entities.

Moreover, Rashid et al. (2019) presented a quality framework for evolving knowledge graphs, which introduces concepts such as persistency (i. e. "characteristics as the degree to which unexpected removal of information from current version may impact stability of the resources." (Rashid et al., 2019)).

While many of these approaches focus on assessing and improving the data after it has been created and published, more and more systems generating RDF programmatically include a quality assessment step in their process, with the purpose of maximizing the quality of the resulting data. That is the case of RML, a relational database to RDF mapping system (Dimou et al., 2015).

2.4 Wikidata

Wikidata is “the free, multilingual knowledge base that anyone can use and edit" (Vrandecic and Krötzsch, 2014). It contains a structured representation of human knowledge, including statements that describe entities and relations between entities in different topic areas ranging from general-audience subjects (e. g. geographical locations and cultural objects), to more specific ones (e. g. genes and citations of scientific publications). As of June 2020, it contains over 87 million items and 7.7 thousand properties.

Wikidata was initially created to support the information management of Wikipedia with a central repository that each individual Wikipedia project (e. g. English, German, Spanish Wikipedias) could consume to preserve data up-to-date across projects, and encourage an homogeneous information coverage. Currently, its data is consumed not only in Wikipedia articles, but also in a wide-variety of systems, including Google’s search engine, which uses Wikidata to cross-check facts in the Google Knowledge Graph, or Eurowing’s flight app, which directly displays Wikidata’s data.

2.4.1 The Data

The data in Wikidata is open [13] and licensed under a Creative Commons public domain license [14], which makes it (re)usable by anyone. While the scope of the data is intended to be very broad, there are some requirements that need to be fulfilled for the data to be accepted in Wikidata. These requirements or *notability criteria* [15] are:

- entity descriptions need to have a link to other Wikimedia projects (e. g. Wikipedia, Wikimedia Commons)
- an entity needs to refer to a “clearly identifiable conceptual or material entity" that “can be described using serious and publicly available resources"
- the entity “fulfills a structural need" that is not covered by other entities

Wikidata plays the role of a secondary source, that points to facts that other external primary sources have published publicly, acknowledging data plurality by design. Hence, there can be two statements describing the same property for the same entity with a different value, because two different external sources may have published contradictory information.

2.4.1.1 Data Model

While Wikidata’s data model resembles RDF in the way statements are constructed, the project’s specific needs led to the definition of their own data model. Data is organized into *items* and *properties*. Items represent either classes of entities (e. g. human, geographic location), or concrete entities (e. g. Douglas Adams, Albert Einstein); while properties represent types of relations between items (e. g. capital of, head of government, instance of), or value-based features (e. g. official name, inception). Each of these items and properties, is identified by a unique identifier with nomenclature `Qnumber` for items, and `Pnumber` for properties.

[13] Open Definition `https://opendefinition.org/`

[14] Creative Commons CC0 License `https://creativecommons.org/share-your-work/public-domain/cc0/`

[15] Wikidata Notability Criteria `https://www.wikidata.org/wiki/Wikidata:Notability`

Figure 2.10 shows the elements in the data model related to items. The *label* provides a human-readable name for the item in a specific language. The *description* is used to clarify the identity of the item and disambiguate in case there are two items with the same name in their label. The *aliases* indicate alternative names for which the item is known (e. g. “Nole" for Novak Djokovic). The rest of the information that describes an item is written in so-called *statements*. These statements, following modeling principles analogous to RDF (see Section 2.1.1), have the shape of `subject predicate object`. Properties are used in the predicate position, while in the object position statements can have literal values, URIs of internal Wikidata items or external entities, quantitative values or the “unknown value" for cases in which such piece of information is not known/available.

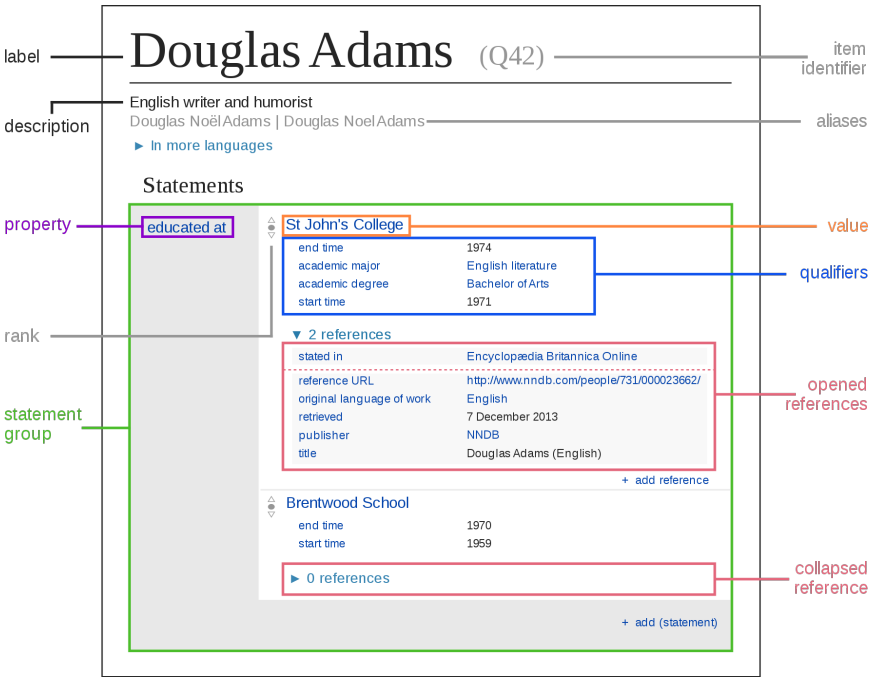

Figure 2.10: Wikidata’s Data Model Visualization (Image created by Charlie Kritschmar (WMDE), under CC0 license)

Statements can be annotated with qualifiers, references and ranks in order to provide more context about the validity and provenance of statements. With qualifiers it is possible to specify, for instance, the period of time in which the value holds (e. g. a person was affiliated with an organization between 2000 and 2010), or the method used to compute the value. References, instead, help specify the primary source that makes the statement verifiable —typically, books, scientific articles or high quality documentation where the piece of information was initially published. When items have multiple values for the same property, ranks can be added to the statement such that the ’preferred’ value is labeled, and deprecation of values is made explicit.

Properties are described in an analogous way, using specific properties for statements (e. g. “`Wikidata property example`", “`subject item of this property`", “`stability of property value`" or “`property constraint`") that describe the way the property is intended to be used. Statements of properties, in contrast to items’ statements, do not need to be referenced.

Recently, Wikidata added *lexemes*, “lexical elements of a language" [16], making Wikidata a valuable source for natural language processing. Lexemes are described indicating their lemma, language, forms, senses (i. e. different meanings) and related statements.

Wikidata and RDF Erxleben et al. (2014) mapped Wikidata to RDF and generated an export mechanism to publish RDF (N-Triples or XML) dumps of the data regularly. Hernández et al. (2015) analyzed alternative representations for data reification and the impact these have in query performance in different SPARQL implementations, finding that the one used in the RDF dumps by Erxleben et al. (2014) was the only one supporting property paths, and identifying no outstanding differences in terms of query performance. Despite this compatibility with RDF, Wikidata statements implement certain Semantic Web constructs differently. For example, in Wikidata items are not typed using the standard `rdf:type` property, but a Wikidata-specific property instead: `wdt:P31` whose meaning is *instance of*. Subsumption is specified via `wdt:P279` instead of `rdfs:subClassOf`, whereas equivalence is indicated via `wdt:P1709` and `wdt:P1628` instead of `rdfs:equivalentClass` and `rdfs:equivalentProperty` respectively. Constraints are defined mainly for data quality assurance purposes, instead of for the inference of new information, as it happens in OWL (Erxleben et al., 2014). The reasons for this situation are primarily practical: the people modeling the knowledge in Wikidata found it easier to define their own predicates. This decision ensures that the meaning of properties (and other schema constructs) have exactly the expected meaning for the environment at hand. The negative aspect of it is that Semantic Web data consumers —who can consume the data via the SPARQL endpoint, URI dereference or using the RDF dumps— need to adjust their queries to include the new (and analogous) properties when querying Wikidata.

2.4.1.2 Data Linking in Wikidata

As a knowledge base complying with the Linked Data principles, Wikidata is connected to many other datasets, such as VIAF, WorldCat, IMDb, and many national museums and libraries. As it happened with many other datasets, integration efforts in Wikidata have focused on the data-level, resulting in a weakly aligned schema but a tight linkage of its items with entities of external datasets. The most frequently type of links implemented in Wikidata corresponds to identity. These so-called authority control links are (in contrast to Semantic Web/Linked best practices) not specified through the `owl:sameAs` predicate. In Wikidata, these links have a dedicated property, one for each dataset that is integrated (e. g. link statements from Wikidata items to VIAF entities are declared with the Wikidata property `wdt:P214 (VIAF ID)`). In total, there are over 4K different (object) properties defined in Wikidata used to point to external entity identifiers [17].

Additionally, Wikidata allows batches of data to be imported directly into the knowledge base. Governmental organizations import their data into Wikidata, not only to complement existing data in Wikidata, but also to increase their data’s reusability (e. g. population statistics gathered by official entities can, this way, be up-to-date in Wikipedia pages). While the provenance of the newly imported statements can (by design) be annotated (e. g. via the `P972` property that points to a catalog instance), many of them lack the information about the original source they were imported from, which makes this kind of data integration implicit.

[16] Wikidata Lexemes `https://www.wikidata.org/wiki/Wikidata:Lexicographical_data/Documentation`

[17] SPARQL query that counts the number of properties used with external identifiers containing "ID" in their English label, executed in the Wikidata Query Service `https://w.wiki/Uua`

The *Wikidata Query Service* allows data consumers to execute federated SPARQL queries using the `SERVICE` construct. However, to ensure high performance, preserve security and guarantee data license compatibility, federated queries can only be executed over the datasets that have been approved and whose federation has been technically enabled. As of June 2020, there are a total of 60 federated SPARQL endpoints [18].

While Semantic Web link discovery tools, such as Silk (Isele and Bizer, 2012) and LIMES (Ngonga Ngomo and Auer, 2011), can be used with Wikidata's (RDF) data, other linking tools tend to be used more frequently for this task, given the context in which data is curated and integrated in Wikidata. OpenRefine is an open source framework for data cleaning and data wrangling that is widely used also in non-Semantic Web environments. It allows users to align CSV data with Wikidata items, making it easy for people to prepare data that can be afterward imported into Wikidata (Carlson and Seely, 2017). Mix'n'Match is an open source tool specifically implemented for Wikidata, that allows users to match records of external catalogs against items in Wikidata [19]. Both OpenRefine and Mix'n'Match provide a highly interactive workflows, in which users can inspect and correct/enhance the algorithms' linking output. Mix'n'Match also publishes the link results in its Web site, so that any Wikidata user (even those who did not initiate a linking workflow) can see them labelled as "matched", "unmatched" or "preliminary matched", and can enhance the machine computation. Both solutions provide batch linking mechanisms that use string comparison rules that may consider entity property values and types.

In the last few years, Wikidata increased its centrality in the LOD cloud. The more its data grew and the more this data was considered and consumed, the more links were created to connect Wikidata's items to other datasets' items, in both directions (i. e. with in- and out-links). It became a so-called interlinking hub (Neubert, 2017), analogously to the role that DBpedia [20] had for the last decade. While both knowledge bases are somewhat related to Wikipedia, they clearly differ in content, purpose and the data management process they adhere to. DBpedia was created as an RDF serialization of the content in Wikipedia's Infoboxes —the tables that contain summarizing facts in Wikipedia articles, whereas Wikidata was created as a stand-alone knowledge base to be consumed by Wikipedia, and whose data increases independently. DBpedia is generated after automatic transformation, while Wikidata's data management workflow heavily depends on human-input, even if it also partly relies on various kinds of automation. Still, DBpedia contains data that Wikidata lacks, and the DBpedia ontology is considerably more stable (as it was curated by knowledge engineering experts). For these reasons, and to support previous DBpedia-related efforts, the DBpedia research community worked on the integration of both knowledge bases (Ismayilov et al., 2018). One of the results of this work, is a detailed mapping between DBpedia and Wikidata entities [21].

[18] SPARQL endpoints federated with Wikidata `https://www.wikidata.org/wiki/Wikidata:SPARQL_federation_input`

[19] Mix'n'Match `https://mix-n-match.toolforge.org/#/`

[20] DBpedia `https://wiki.dbpedia.org/`

[21] Mappings between DBpedia and Wikidata entities `https://wiki.dbpedia.org/get-involved/google-summer-code/gsoc-2013/ideas/wikidata-mappings`

2.4.2 A Community-Based Knowledge Base

A characteristic that distinguishes Wikidata from other knowledge bases, such as YAGO or DBpedia, is that it is curated and maintained by a large community of volunteers. These people are typically intrinsically motivated, and committed to "giving more access to more knowledge" [22]. A study by Nov (2007) surveyed Wikipedia editors in order to identify the reasons behind their non-paid editing work. The 151 valid responses collected indicated that Wikipedia editors contribute primarily for fun; other reasons followed in order of importance were: "ideology and values", understanding other perspectives and the feeling of being needed by others, and social effects such as "feeling less lonely", establishing useful contacts or being encouraged to write in Wikipedia by a close network of friends or family members. While (to the best of my knowledge) an analogous and more recent survey has not been conducted among Wikidata editors, it is legitimate to assume that Wikidata editors follow a very similar (if not the same) motivation scheme, as all the projects in the Wikimedia movement share a vision, code of conduct, as well as a generic infrastructure and social processes.

Let us revise the way the Wikidata community works, as well as the data quality management process in place, and the interplay between human and machine computation.

[22] Wikidata presentation by Lydia Pintscher at Wikimania 2015 https://wikimania2015.wikimedia.org/wiki/Submissions/State_of_Wikidata_-_giving_more_people_more_access_to_more_knowledge_one_edit_at_a_time

2.4.2.1 Peer-Production System

Wikidata is a peer-production system, which according to Benkler et al. (2015) can be defined as "a form of open creation and sharing performed by groups online that: set and execute goals in a decentralized manner; harness a diverse range of participant motivations, particularly non-monetary motivations; and separate governance and management relations from exclusive forms of property and relational contracts (i.e., projects are governed as open commons or common property regimes and organizational governance utilizes combinations of participatory, meritocratic and charismatic, rather than proprietary or contractual, models)".

Deliberation In this peer-production system, the community defines best practices and makes decisions as a collective. The basis of this collective and cooperative action is deliberation. Wikidata community members constantly discuss needs, requests, proposals, plans and already implemented actions, that may relate to data and schema modeling, as well as software infrastructure and social rules. These discussions, which are documented online (mostly within the Wikimedia space, but also in complementary channels), materialize in different ways: for instance, every page in the wiki —be it an item, a property or a project page— has a discussion or talk page associated, so as to allow community members to raise and comment on issues that refer to the item/property/project at hand. Additionally, there are special pages where specific proposals are collected and discussed until the community has reached consensus and/or a decision has been defined. For instance, the creation of new properties in the schema [23], and the federation of SPARQL endpoints [24] are proposed this way, to allow community members to express and argue their *support* or *rejection* to the proposed ideas. For software development, the core Wikimedia engineering team uses a ticketing system (Phabricator [25]) to coordinate project management, in which community members may submit tasks for reporting not only bugs, but also for requesting new features that are needed.

In so-called *Requests for Comments* (RfC) [26] —a procedure inherited from the ARPANET project and used by many Task Force actions on the Web, community members can request the feedback from the community on a concrete topic that requires broader discussion. For example, there have been RfCs to discuss specific topics such as merging two items, and how to model personal names, but also more general topics such as the definition of data quality in Wikidata or the notability criteria. Ideally, a valid RFC results in an actionable outcome, after reaching consensus for or against a proposal. However, RfCs may become closed as inconclusive or with a lack of consensus (i.e., when the community considers that there has been sufficient discussion but there is disagreement), or declared stale (i.e., the discussion becomes inactive for a long period of time).

The CSCW literature has studied online deliberative systems, identifying human factors that influence the process. Im et al. (2018) studied RFCs in Wikipedia and identified that some of the most common reasons for RfCs to become stale are: (i) issues with the initial proposal (e. g. it has an unclear or biased formulation), (ii) "bickering and sock-puppeting", (iii) lack of expertise and interest of the discussion participants, (iv) the discussion seems to have an obvious consensus and people interpreted that it did not need closure, (v) over-complexity of the RfC and (vi) political reasons among community members. Laniado et al. (2012) analyzed the emotions shown by Wikipedia editors in community

[23] Property Proposal Page in Wikidata `https://www.wikidata.org/wiki/Wikidata:Property_proposal`
[24] Wikidata Federation Input `https://www.wikidata.org/wiki/Wikidata:SPARQL_federation_input`
[25] Wikidata Phabricator `https://phabricator.wikimedia.org/project/view/71/`
[26] Wikidata Requests For Comments `https://www.wikidata.org/wiki/Wikidata:Requests_for_comment`

discussions and concluded that “women tend to participate in discussions with a more positive tone”. Schaekermann et al. (2018) implemented a deliberation interface to study factors that influence the resolvability of deliberation in two MTurk crowd tasks (i. e. a sarcasm classification task and a relation validation task). Their research pointed to "the initial disagreement", the volume and quality of the deliberation, and the features of the task at hand as influencing factors for resolvability. Whether these findings directly apply to the Wikidata process remains an open research question.

Goal-Oriented and Self-Organized Cooperative Work While every Wikidata editor may freely decide what to edit and the means to implement the edit(s), more experienced editors have demonstrated the effectiveness of goal-oriented and organized cooperative work. Not only does this editing strategy give a sense of meaning and cohesion to editors, but it also helps to achieve large volumes of high quality data. These organized data curation efforts are defined around *Wikiprojects* [27], which aggregate community members around a specific focus (e. g. movies, research data, culture, history, citations, etc.). Kanke (2019) investigated the “activities, tools and norms" in 5 Wikidata Wikiprojects, manually processing the discussion pages and using the Activity Theory (Kaptelinin and Nardi, 2012) framework as a basis. The author classified the cooperative work into activities such as “conceptualizing the curation process", “appraising objects", “ingesting objects from external sources", “creating or re-organizing collaborative infrastructure", and “welcoming newcomers". The tools used by editors throughout these activities range from scripts that help generate lists automatically and bots that automate and batch edits, to tools that have been implemented to cover a specific use case, like the Primary Sources tool that enables an editor-oriented reviewing workflow of imported data that was initially used in the context of the Freebase import. In terms of regulating these activities, the author identified that “it includes assessing acceptable editor conduct and creating a structure to train beginners in how" (i.e. where this should be implemented and the way comments should be elaborated) “the community prefers contributions".

Besides Wikiprojects, editors also focus their editing attention around topics when they participate in competitions initiated the Wikimedia chapters (e.g., ensuring that all museums in a region are described in Wikidata, or guaranteeing that Wikidata items have pictures in Commons). The “100wikidays"initiative, applicable to Wikidata, focuses on creating an editing habit rather than a topic: it was created as a challenge by a Wikimedian to encourage editors to edit 100 consecutive days.

Müller-Birn et al. (2015) studied emerging editor task-oriented roles through an empirical analysis of editing patterns in Wikidata, in the early stages of the knowledge base (from December 2012 to October 2014, when the size of the data was around 17 million items). The authors looked at edits done in item and property spaces, distinguishing between human registered editors, non-registered editors, and bots. After applying a clustering algorithm, the roles identified are: item creator (who creates new items), item expert (who adds many statements to describe the item), item editor (who edits some values in the statements, adds sitelinks), property engineer (who creates properties and edits their talk pages), property editor (who ads property statements) and reference editor (who adds sitelinks to connect the items to Wikipedia or other Wikimedia projects). As the authors acknowledged, the results “suggest very specialised contributions from a majority of users. Only a minority, which is the most active group, participate all over the project" (Müller-Birn et al., 2015).

Transparency, Accountability and Recognition Every single change (or edit) implemented in the knowledge base is visible, not only to the community but to the whole world. Wikidata publishes

[27] Wikiprojects `https://www.wikidata.org/wiki/Wikidata:WikiProjects`

for every editing action, the identifier of user who implemented it, the timestamp when the action occurred, the element edited (i. e. item/property/talk/project/user page), the type of action implemented (i. e. added/changed/removed a certain piece information) and other annotations contextualizing the edit that may be manually provided by the user, or automatically flagged by the Wikidata systems.

Human users can edit Wikidata in different ways: as registered users (after creating a user account in Wikidata), or as unregistered users. In both cases anonymity may be preserved, as a Wikidata user account only requires an email address and username, which may or may not reveal the real-world identity of the person behind the user. Edits implemented by unregistered users are signed with their IP address. By registering, other community members can address the user, and their edits can more easily be monitored, in case of conflict. Hence, registering as a user can expedite healthy cooperation.

This transparency is fundamental for accountability, and it generally contributes to regulating user behavior in the community. Moreover, it facilitates recognition for contribution. The community maintains statistics about the most active editors, which rather than being a trigger for unhealthy competition, aid value others' work. Additionally, editors can send each other a message of gratitude and thank each other for a concrete edit.

User groups The Wikidata community, analogously to any other Wikimedia subcommunity, is comprised of members with different roles. This distribution of tasks and responsibilities structures cooperation in the community.

The staff employed at Wikimedia Deutschland —the Wikimedia chapter that created and runs Wikidata— contributes with a myriad of tasks that range from product and project, to software engineering and system administration to ensure that the infrastructure serves the needs of the community. However, as an open source project, the core infrastructure is extended by community members who develop and share software for more specific purposes, as explained in the next section. Community growth efforts are scattered over the world, involving local groups that co-operate with the different Wikimedia chapters (e. g. Wikimedia Switzerland, Wikimedia Spain, Wikimedia Foundation etc.) combining staff and volunteers.

Based on experience and purpose, editors are assigned to user groups with different permissions [28]. The status of user accounts is public and can be consulted (manually and programmatically) at any point in time. When users register they are automatically classified as *new*, until they accumulate "at least 4 days of (Wikidata) age and at least 50 edits", which is when they are updated to *autoconfirmed users*. With this transition, users no longer need to answer to CAPTCHAs that try to discard robots and can implement new actions, such as re-locating some type of Wikidata pages. These users can create and update items, but cannot, for example, create properties or delete any data from Wikidata. These are special rights reserved for users who have already exhibited good faith behavior, commitment and value through (usually) a large volume of edits.

To be manually classified into further user groups that provide such permissions in this trust-based system, a user can post a well-documented *Request for Permission* [29], and similarly to RfCs, the community discusses the case and collectively decides to approve or reject that the requested permission is granted. There are various user groups for these special rights: *administrators* are the editors with one of the highest levels of access and may block users, create properties, close RfCs and delete (item) pages, among other things; *bureaucrats* can assign and revoke permissions to users, switching their user group; *translation administrators* can have admin access to the software extension used for translating

[28] Wikidata User Levels `https://www.wikidata.org/wiki/Wikidata:User_access_levels`
[29] Request for Permissions `https://www.wikidata.org/wiki/Wikidata:Requests_for_permissions`

content in many different natural languages; *bots* can execute a large volume of edits in an automated way; and *flooders* can implement manually (as humans) repetitive edits.

Moreover, there are specific permissions that can be assigned to users to, for instance, rollback an edit, create properties, edit from an IP that has been blocked, or become an ad-hoc confirmed user. These user groups have noticeboards, where other users may request their attention and input.

Additionally, Wikimedia communities and the Wikimedia Foundation elect people that have roles to check users, protect the content and ensure its quality —that is the case of *stewards* and *ombudsmen*.

2.4.2.2 Humans and Machines in Wikidata

While Wikidata is evidently a human-centered knowledge base, where decisions and actions by human users supersede those by machines, the ever-growing scale of the data demands a hybrid ecosystem, in which machines help humans automating repetitive data management actions. Based on the way they operate, three different types of algorithmic data management solutions can distinguished in Wikidata:

1. **Tools and Scripts**: software with or without a front-end that may automatically read or edit Wikidata's data, and whose purpose is to provide users (editors or consumers) with a concrete functionality that may help their work [30]. Examples of these tools include tools for visualizing Wikidata's data (e. g. Histropedia showing timelines), data monitoring tools (e. g. Property Explorer to explore the list of existing properties, Lonely Items to browse the set of items that are not highly linked), and tools that facilitate editing that solves a concrete need (e. g. Wikidata Game asks to validate the gender of a person item, or the colour of a flower), or one-time bulk editing (e. g. QuickStatements allows users to import CSV data into Wikidata). Edits done by tools are attributed to the logged-in user using the tool. Some of the tools, such as QuickStatements, leave a trace in the edit (or revision) comment in order to be recognizable.

2. **Bots**: software that runs constantly and independently and can read and edit high volumes of data [31]. Due to the substantial impact bots may have, the community requires bots' conduct to be reviewed before they can run with the corresponding privileges (i. e. at a higher edit rate). Bots need to make a request for approval, showcase their automatic behavior for 50-250 edits and if they obtain support by the community, the request is closed by an administrator and the bot receives a bot flag by a bureaucrat (i. e. the user account of the bot is assigned to the bots user group). Examples of bots include DBpedia-mapper-bot, that automatically creates mappings between Wikidata and DBpedia entities; GitHub-wiki-bot, that collects and imports metadata from GitHub about software projects that appear in Wikidata; and the Elhuyar Fundazioa bot, which imports Basque lexemes, their forms and definitions from the Elhuyar Student Dictionary. Edits implemented by a bot are attributed to the bot's user account, which is publicly linked to the owner's user account.

3. **Hidden tools**: software that automates edits without disclosing its computational nature. The Wikidata API exposes functionality to edit Wikidata programmatically, without logging in. Hence, there might be tools behind IPs. Recently, Hall et al. (2018) created a predictive model that given a set of behavioral traces classifies a user as bot or non-bot. Their supervised method is able to perform with high fitness ("PR-AUC: 0.845, ROC-AUC: 0.985"), analyzing features

[30] Wikidata Tools `https://www.wikidata.org/wiki/Wikidata:Tools`
[31] Wikidata Bots `https://www.wikidata.org/wiki/Wikidata:Bots`

that describe patterns in the activity (e. g. number of edits in each namespace (items/properties/users), the mean time between edits) and the content of the activity (e. g. “number of unique claims changed", words appearing in the comments of the edits whose ending contains “bot"). The authors applied the model to the dataset containing edits by non-bot registered users and unregistered users, between November 2012 and April 2017, and identified that not only “2% of anonymous “human" user edits" (unregistered users) “overall are from bots", but also “3% of registered “human" user edits overall".

Bots perform a very large part of the total edits in Wikidata. Statistics provided by Wikimedia dashboards (see Figure 2.11 [32]) show that depending on the time window, bot edits represent half of the edits done overall. Figure 2.11 also indicates a considerable increase in the volume of edits done by registered users, which can be explained by the emergence of tools for importing, linking and extending descriptions of items (in batches). The study by Müller-Birn et al. (2015), showed that (in the first two years of Wikidata) the distribution of tasks for each of the identified user roles was analogous in registered users and bots.

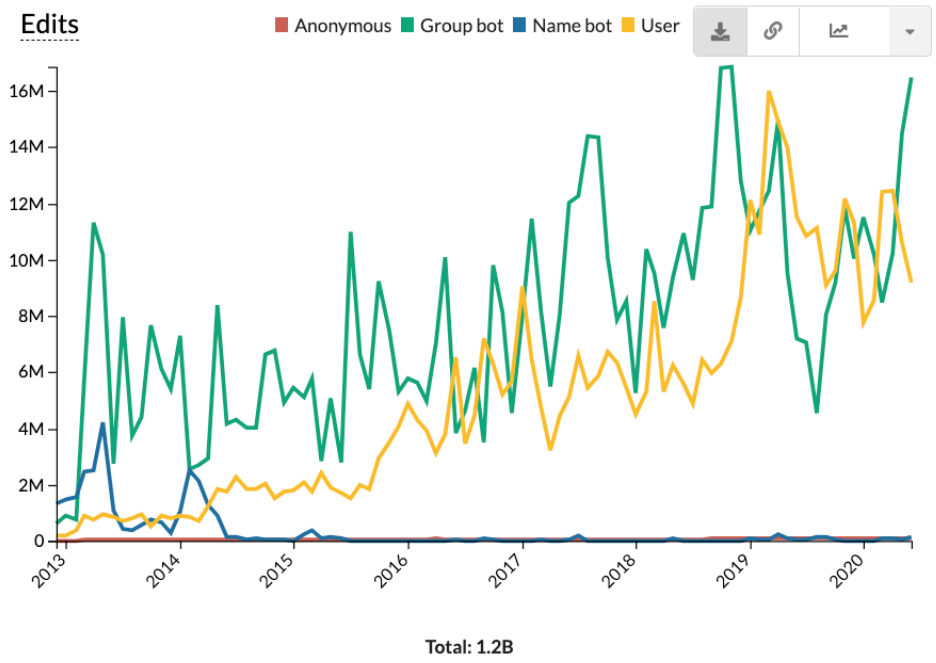

Figure 2.11: Total count of edits in Wikidata per type of user. The orange line shows edits by unregistered users, the yellow line shows edits by registered users who are declared as human users, the blue line shows edits by users whose username contain the string “bot", and the green line shows edits by users that haven been assigned to the bots user group (i. e. received the bot flag). Source: Wikimedia

Despite their autonomy while executing their edits, bots are part of the social ecosystem in Wikidata. They are implemented, administered and supervised by humans, and can, consequently, be considered as an extension of their owners’ activity. Geiger and Halfaker (2017) acknowledged this consideration

[32] Wikidata Statistics by Wikimedia `https://stats.wikimedia.org/`

in the context of Wikipedia, when they replicated and extended the work by Tsvetkova et al. (2017) on the interaction of bots. The initial study characterized edit conflicts between bots, by analyzing the edits by bots reverted by other bots. The work by Tsvetkova et al. (2017) found, for example, that "over the ten-year period, bots on English Wikipedia reverted another bot on average 105 times, which is significantly larger than the average of 3 times for humans" and highlighted the fact that bots edit with a lack of control and governance. In their later study, and following a trace ethnography methodology, Geiger and Halfaker (2017) re-analyzed the data, mining the reverts together with contextual information of the edits, and taking other editing dynamics into consideration (e. g. the way pages are moved or renamed by the community). This new and deeper study concluded that "the majority of bot-bot reverts on articles are better characterized as routine, productive, and even collaborative work between works". Through their analysis, Geiger and Halfaker (2017), who additionally to their scientific background are experienced Wikipedia community members, shed light on the nature of the bot-bot reverts: sometimes the origin of these lay in a bug in the implementation, that once noticed by the community was corrected by the bot owner, otherwise when bot owners become unresponsive their bots tend to be blocked. Some other times they are the result of conflicts between Wikipedia community members, either because they materialize different processes for a similar task or because they have different conceptions of what should be done in Wikipedia. All in all —the authors indicated— "approximately 0.5% of bot-bot reverts to articles constitute bot-bot conflict" (i. e. bots have an opposing behavior).

Farda-Sarbas et al. (2019b) studied the RfPs posted by bot owners to the Wikidata community, identifying that only 12% out of 681 requests they revised (from November 2012 until July 2018) are closed as unsuccesful; the remaining 600 were declared successful by the community, leading to 391 different bots being approved to operate in Wikidata. The authors observed that from the tasks that each of these requests addressed, the most prominent one was to add data into Wikidata. Additionally, this study encoded the reasons for unsucessful requests, concluding that the most frequent reasons are: (i) the owner becomes unresponsive, (ii) the owner of the bot withdraws the request, (iii) the request is redundant with respect to others, (iv) the community does not manage to reach consensus, (v) the community does not trust the owner of the bot after undesirable previous behaviour, (vi) the proposed task does not need a bot, or a bot flag, (vii) there is a lack of information and (viii) the bot is trying to do something that goes against Wikidata policies. These findings show a healthy co-existence of bots in the Wikidata community.

2.4.2.3 Data Quality Management in Wikidata

The central element to data quality management in Wikidata is the collaborative human effort that entails observing changes occurring in the knowledge base and improving anything that can be improved. As mentioned before, machines (tools and bots) contribute by adding and updating data, but their operations are carefully supervised by the community of humans, before and while they run independently.

For this kind of monitoring, the Wikidata community has several procedures and tools that highlight changes ranked by recency. Some of these tools have a general scope. For example, Recent Changes [33] and the reCh tool [34] list single recent writing actions that were implemented in Wikidata's data. Speed patrolling [35] also shows recent changes, together with a comparison between two versions of the

[33] MediaWiki Recent Changes Feature for Wikidata `https://www.wikidata.org/wiki/Special:RecentChanges`

[34] reCh `https://pltools.toolforge.org/rech/`

[35] Speed Patrolling Tool `https://www.wikidata.org/wiki/User:Lucas_Werkmeister/SpeedPatrolling`

data, generated before and after the monitored action was implemented, such that users can act upon changes, by either rolling them back or by labeling them as patrolled (i. e. supervised by a human). Editors using the wiki may also add specific items to a watchlist, in order to receive notifications when changes occur in their page.

The data quality dimensions enumerated in both the framework by Wang and Strong (1996) and its Linked Data extension by Zaveri et al. (2016) are, in principle, relevant to Wikidata. However, Wikidata's technical characteristics and community dynamics set new requirements for measuring these standard data quality dimensions. For instance, as noted by one community member during an RfC to discuss the meaning and relevance of state-of-the-art data quality frameworks (Piscopo and Community, 2016), given the lack of a "standard class system" it seems desirable to measure schema completeness in terms of properties exclusively, even if schema completeness usually considers classes as well. Moreover, when measuring semantic accuracy and consistency, one needs to acknowledge that Wikidata is a secondary source, representing information that other primary sources publish, and thus, may contain different values for properties that are expected to have one single value associated (e. g. date of birth or place of birth of a person).

A recent literature review (Piscopo and Simperl, 2019) concluded that most of the scientific work around Wikidata's data quality management methods focused on the *completeness* dimension. This finding concurs with the result that we obtained after labeling available Wikidata tools in terms of the data quality dimension they primarily address (Farda-Sarbas et al., 2019a). It seems reasonable that during the first years of the project, completeness received more attention than other dimensions, as monitoring completeness encourages the growth of the knowledge base, especially since accuracy management is addressed by the community. However, the larger the data and the larger the variety of people involved is, the higher the need is for identifying disagreement in interpretations, and inaccurately imported or referenced data. Two examples of tools that help editors improve Wikidata in terms of completeness are ReCoin and Integraality [36]. ReCoin is implemented as a script that editors can activate within the wiki, in order to see a relative completeness assessment for a given item —it indicates how well described the item is, compared to other items of the same type, by observing its statements' properties. Integraality builds a dashboard that helps to monitor the presence of a self-defined set of properties in a self-defined set of types of items. It is thus useful to identify missing data in items that are relevant to a certain topic, that may be related to, for example, Wikiprojects.

Besides the provision of technological infrastructure, the community makes an effort to identify and disseminate best practices in order to improve the quality of the data. To this end, community members selected a set of items as *showcase items* —item descriptions that serve as a reference for anyone looking for examples of how to edit, for example, a person/place/book's description in Wikidata [37].

Piscopo et al. (2017) analyzed item co-editing in order to understand the kind of groups of editors that are related to high-quality item editing. Their study observed the edit history of a set of 5K items whose quality had been labeled by an expert user in a previous work. Item quality was defined as a multi-label scheme (A-E) that defines the extent of "expected" information present in the item description. For example, items of class A contain "all relevant statements, with solid references, and complete translations, aliases, sitelinks, and a high quality image", whereas items of class C contain "most critical statements, with some references, translations, aliases, and sitelinks", and items of class D contain "some basic statements, but lacking in references, translations, and aliases" (Community, 2017; Yapinus et al., 2017). After applying a regression model with independent variables including tenure and

[36] Wikidata Tools `https://www.wikidata.org/wiki/Wikidata:Tools`
[37] Showcase Items in Wikidata `https://www.wikidata.org/wiki/Wikidata:Showcase_items`

interest diversity, as well as the proportion of bots and anonymous edits, the authors concluded that the proportion of bots positively influences the resulting quality of items, while the proportion of edits done by unregistered editors has a negative influence on the resulting item quality. Tenure diversity and interest diversity also showed a positive influence in item quality, although tenure diversity (measured as a variation coefficient in lifespan) was more prominent than interest. Kaffee et al. (2019) studied humans and bot editing patterns in multi-linguality management in Wikidata, observing edits implemented in item labels. Their results show a clear differentiation between humans' and bots' activity in terms of the number of the natural languages they edit in: the majority of registered (around 63%) and unregistered (around 87%) editors edit in one natural language, whereas half of the bots edit in one language, and the other half implement changes in multiple languages (around 37% between 2 and 5 languages, and the rest even more).

Chapter 3

CrowdMap: Crowdsourcing Ontology Alignment with Microtasks

Abstract: The last decade of research in ontology alignment has brought a variety of computational techniques to discover correspondences between ontologies. While the accuracy of automatic approaches has continuously improved, human contributions remain a key ingredient of the process: this input serves as a valuable source of domain knowledge that is used to train the algorithms and to validate and augment automatically computed alignments. In this paper, we introduce CrowdMap, a model to acquire such human contributions via microtask crowdsourcing. For a given pair of ontologies, CrowdMap translates the alignment problem into microtasks that address individual alignment questions, publishes the microtasks on an online labor market, and evaluates the quality of the results obtained from the crowd. We evaluated the current implementation of CrowdMap in a series of experiments using ontologies and reference alignments from the Ontology Alignment Evaluation Initiative and the crowdsourcing platform CrowdFlower. The experiments demonstrated that the overall approach is feasible, and can improve the accuracy of existing ontology alignment solutions in a fast, scalable, and cost-effective manner.

3.1 Introduction

The last decade of research on ontology alignment has brought a wide variety of automatic methods and techniques to discover correspondences between ontologies. Researchers have studied extensively the strengths and weaknesses of existing solutions, as well as their natural limitations and principled combinations, not least through community projects such as the Ontology Alignment Evaluation Initiative (OAEI).[1] Partly as a result of these efforts the performance of the underlying algorithms has continuously improved. However, most researchers believe that human assistance is nevertheless required, even if it is just for the validation of automatically computed mappings. In this paper, we introduce CrowdMap an approach to integrate human and computational intelligence in ontology alignment tasks via microtask crowdsourcing.

The term "microtask crowdsourcing" refers to a problem-solving model in which a problem is outsourced to a distributed group of people by splitting the problem space into smaller sub-problems,

[1] http://oaei.ontologymatching.org/

or tasks, that multiple workers address independently in return for a (financial) reward. Probably the most popular online instantiation of this model is Amazon's Mechanical Turk (MTurk) platform (`https://www.mturk.com/`) which offers a virtual labor marketplace for microtasks as well as basic services for task design and publication, work assignment, and payment. Typical problems that are amenable to microtask crowdsourcing are those problems that we can easily distribute into a (high) number of simple tasks, which workers can complete in parallel, in a relatively short period of time (in the range of seconds to minutes), and without specific skills or expertise. Examples of such problems include finding a specific piece of information on the Web, labeling or classifying content, and ranking a list of objects. Recently, researchers have demonstrated the effectiveness of microtask crowdsourcing for far more complex problems by using sophisticated workflow management techniques on top of the basic services of existing platforms, and optimizing quality assurance and work assignment (Ipeirotis et al., 2010; Kulkarni et al., 2011; G. Little and Miller, 2009). As a result, microtask crowdsourcing has been successfully applied to a broad range of diverse problems: completing surveys, translating text from one language to another, creating comprehensive product descriptions, matching pictures of people, summarizing text (Bernstein et al., 2010) and many others.

Ontology alignment is a good fit for microtask crowdsourcing for several reasons. First, verifying whether or not a mapping is a correct one is naturally a microtask, and workers do not need much context to figure out the right answer. Second, we can easily decompose the overall problem of verification of a set of candidate mappings into atomic tasks corresponding to the individual mappings. These tasks are largely independent of one another. Third, while ontologies can be quite large (with tens of thousands of classes), their scale is often considerably smaller than the scale of the data itself. Thus, crowdsourcing becomes a tractable way to verify all candidate alignments between two ontologies. Finally, ontology alignment is still one of those problems that we cannot automate completely, and having a human in the loop might increase the quality of the results of machine-driven approaches.

There are two different ends of the spectrum in which we envision applying crowdsourcing to ontology alignment. On the one hand, we can generate all possible pairs of alignments between two ontologies, and ask the crowd to evaluate each of the candidates. However, this option will clearly not scale well, as we will be asking the users to inspect an extremely large number of pairs—equivalent to the cartesian product of the size of the two ontologies—and we know that the number of valid correspondences are usually at most comparable to the number of terms in the smaller of the two ontologies. On the other hand, we can start by running an automatic algorithm that generates potential alignments, and subsequently have the crowd assess the results. This second option will likely be much more scalable in terms of the number of tasks and answers needed from the crowd (and thus the duration and cost of the alignment exercise). While this scenario is likely to lead to improvements in the precision of the original algorithm, with this approach we will be able to have similar effects also on the recall if we present the crowd with the very low confidence mappings.

CrowdMap is a new model for ontology alignment which uses microtask crowdsourcing to improve the accuracy of existing automatic solutions. In evaluating this approach, we explore the following research questions:

R1 Is ontology alignment amenable to microtask crowdsourcing?

R2 How does such a human-driven approach compare with automatic (or semi-automatic) methods and techniques, and can it improve their results?

R3 What types of alignment problems can workers feasibly solve? What correspondences between

elements of different ontologies (e.g., similar, more general, more specific) can be reliably identified via crowdsourcing?

We introduce CrowdMap and its implementation using CrowdFlower (`http://crowdflower.com/`) a crowdsourcing platform which acts as an intermediary to a number of online labor marketplaces, including MTurk. For a given pair of ontologies, CrowdMap translates the alignment problem into microtasks that address individual alignment questions, publishes the microtasks on an online labor market, and evaluates the quality of the results obtained from the crowd. We tested the current implementation in multiple settings in order to determine how we can optimize the quality of the crowdsourced results through specific task-design and work-assignment features. For this purpose we ran a series of different experiments: an exhaustive alignment between two (smaller) ontologies; a broader set of ontologies assessing the outcomes produced by a simulated automatic algorithm; and validating the mappings computed by one of the algorithms that participated in Ontology Alignment Evaluation Initiative. The experiments provided evidence that the overall idea to apply microtask crowdsourcing to ontology alignment is not only feasible, but can also significantly improve the precision of existing ontology alignment solutions in a fast, scalable, and cost-effective manner. The findings of the experiments allowed us to define a number of best practices for designing purposeful ontology alignment projects, in which human and computational intelligence are smoothly interwoven and yield better results in terms of costs and quality compared to state-of-the-art automatic or semi-automatic approaches.

3.2 Related Work

While the ontology alignment community acknowledges the importance of human contributions, the question of how to optimally collect and harvest these contributions leaves room for further research (Shi et al., 2009). Falconer and colleagues described the results of an observational study of the problems users experience when aligning ontologies (Falconer and Storey, 2007). They emphasized the difficulties experienced by laymen in understanding and following the individual steps of an alignment algorithm. In our work, we provide further evidence for the extent to which contributions from non-technical users can provide valuable input in the alignment process, and investigate alternative means to describe and document alignment tasks in order to make them accessible to laymen.

Another approach employs Web 2.0 technologies and principles to engage a community of practice in defining alignments, thus increasing the acceptance of the results, and reducing or distributing the associated labor costs (McCann et al., 2008; Hausenblas et al., 2009; Noy et al., 2008; Zhdanova and Shvaiko, 2006). An early proposal on collaborative ontology alignment by Zhdanova and Shvaiko (2006) developed a community-driven service that allowed users to share alignments in a publicly available repository. BioPortal (Whetzel et al., 2011) offers a comprehensive solution in the biomedical domain. It enables users to create alignments between individual elements of an ontology (Noy et al., 2008). However, in these approaches, the solicitation for the mappings is "passive": the users must come to the site, find the terms of interest, and create the mappings. There is no expected reward, other than community recognition. By contrast, our CrowdMap model is essentially "mapping for hire" where we do not expect users to have a specific interest in the task that they perform other than the monetary reward that they get. Our experience shows that there is no comparison in the quantity of the work that can be obtained via volunteering and microtask crowdsourcing: putting aside the different knowledge domains that the two approaches address, we were able to get orders of magnitude more alignments in a day in the experiments with the current CrowdMap implementation than BioPortal

received in a year. In this paper, we evaluate the quality of these mappings to determine how useful the microtask-based alternative is beyond the actual number of mappings generated.

McCann and colleagues studied motivators and incentives in ontology alignment (McCann et al., 2008). They investigated a combination of volunteer and paid user involvement to validate automatically generated alignments formulated as natural-language questions. While this proposal shares many commonalities with CrowdMap, the evaluation of their solution is based on a much more constrained experiment that did not rely on a real-world labor marketplace and associated work force.

Games with a purpose, which capitalize on entertainment, intellectual challenge, competition, and reputation, offer another mechanism to engage with a broad user base. In the field of semantic technologies, the OntoGame series proposes several games that deal with the task of data interlinking, be that in its ontology alignment instance (SpotTheLink (Thaler et al., 2011b)) or multimedia interlinking (SeaFish (Thaler et al., 2011a)). Similar ideas are implemented in GuessWhat?!, a selection-agreement game which uses URIs from DBpedia, Freebase and OpenCyc as input to the interlinking process (Markotschi and Völker, 2010). While OntoGame looks into game mechanics and game narratives and their applicability to finding similar entities and other types of correspondences, our research studies an alternative crowdsourcing strategy that is based on financial rewards in a microtask platform.

More recently, researchers in the Semantic Web community have begun to explore the feasibility of crowdsourcing for assigning URIs to entities that are discovered in textual Web pages. ZenCrowd, for example, combines the results of automatically and human-generated answers to link entities recognized in a text with entities in the Linked Open Data cloud (Demartini et al., 2012). ZenCrowd developers proposed a variety of techniques to reduce the scope of the crowdsourcing task, such as excluding candidates for which an algorithm already has a high confidence score from the set to be validated. Our approaches are similar in spirit (using the crowd to improve the performance of automatic algorithm in alignment). However, ontology alignment (rather than data alignment) has a more tractable scope. The motivation of our work is also different: our goal is not to identify which of the two approaches (machine vs human-driven) are likely to be more reliable, but to enhance the results produced by an automatic algorithm.

3.3 The CrowdMap Definition and Implementation

CrowdMap takes as input a set of *candidate mappings* between two ontologies and uses a *microtask platform* to improve their accuracy. The model is not bound to a specific instantiation of microtask platform. It can be applied to any virtual labor marketplace that enables requesters to post a problem as a set of independent *microtasks*, which are performed in parallel by *workers* in return for a (usually monetary) reward. In fact, we can apply the same model to other approaches to human computation, such as games with a purpose, which, though operating on different motivational factors, address similar types of problems: decomposable, verifiable, and not requiring domain-specific knowledge or skills.

3.3.1 Fundamentals of Microtask Crowdsourcing

In order to use a microtask platform, a requester packages the work into microtasks and publishes them in batches or groups. Amazon Mechanical Turk (MTurk), one of the most popular crowdsourcing platforms, refers to microtasks as *Human Intelligence Tasks (HITs)*, a term that we will use interchangeably with microtask.

A requester specifies a number of configuration parameters such as the number of answers that she needs for each HIT, the time to complete a HIT, and restrictions on the profile of the workers (e.g.,

geographical location, knowledge of a specific natural language). As most HITs can be solved quickly (within seconds or minutes at most), similar HITs are typically organized into groups or batches which share the same configuration parameters; workers prefer to be assigned to such larger chunks of work instead of dealing with atomic questions in separate processes. Upon completion of the tasks by workers, the requester collects and assesses the responses and rewards the accepted ones according to the pre-defined remuneration scheme. For most platforms, the requester can automate the interaction with the system via an API, while the workers undertake their tasks using a Web-based interface generated by the requester. The overall effectiveness of crowdsourcing can be influenced dramatically by the way that the requester packages a given problem as a series of microtasks (Kittur et al., 2008; Franklin et al., 2011). This packaging includes, in particular, the design of the interface (including clear instructions for the completion of the task, minimal quality criteria for the work to be accepted, and purposeful layout), and the procedures that the requester uses in order to evaluate the results and to measure the performance of workers. Because multiple workers can perform the same microtask, the requester can implement different types of quality assurance (Ipeirotis et al., 2010). For example, one can use majority voting (take the solution on which the majority of workers agree), or more sophisticated techniques that take into account, for instance, the (estimated) expertise of specific workers, or the probabilistic distribution of accuracy of the answers of a given worker. In addition, the requester needs to implement mechanisms to avoid and detect spam in order to reduce the overhead associated with the evaluation of the crowd-produced results. Other factors that are proven to influence the success of crowdsourcing (in particular in terms of the duration of the execution of the tasks, and the ability to find appropriate work resources in due time) are the number of HITs per batch, and the frequency of publication of similar HITs groups, and the novelty of the tasks. Studies showed that whereas grouping HITs into batches leads to economies of scale, batches of several hundreds of HITs are more difficult to assign than the ones with a size up to 100 questions (Franklin et al., 2011). An analogously motivated behavior of workers tending to focus their resources on similarly scoped tasks makes finding assignments for larger problems divided into several batches and HITs more challenging, as finding different eligible workers in due time to address the entire body of work becomes more difficult. Researchers have studied ways to expand the original application scope of MTurk and alike to more complex workflows (G. Little and Miller, 2009), problems with an open, unknown set of solutions (Bernstein et al., 2010), or those characterized by tight time-to-completion constraints Bernstein et al. (2012).

CrowdMap uses CrowdFlower, one of the leading crowdsourcing platforms as a basis for its implementation. CrowdFlower is an intermediary: it is not itself an online labor market, but it publishes microtasks to different crowds simultaneously (including MTurk, Crowd Guru, getpaid, Snapvertise, and others). It implements advanced quality assurance methods based on golden standards in addition to the basic functionality of the crowdsourcing platforms that it accesses. Specifically, CrowdFlower uses "golden units" to denote those types of alignment questions, for which the answer is trivial or known in advance. CrowdMap evaluates whether or not a worker can be trusted by extrapolating from the accuracy of the answers she gave to these particular questions. These methods help determine the reliability and performance of workers, and to filter spammers at run time (Oleson et al., 2011). The terminology used by CrowdFlower to denominate the core concepts of microtask crowdsourcing is slightly different than the one adopted by MTurk. HITs or microtasks are termed "jobs", and answers (or "assignments" in MTurk) to these questions are "judgements". HITs become "job assignments" in CrowdFlower when they are built using job templates and particular data "units". MTurk organizes CrowdFlower tasks in "batches", when they share the title. In the remainder of the paper we will use these terms interchangeably.

3.3.2 The CrowdMap Workflow

The CrowdMap task is to find a set of mappings between two ontologies, O_1 and O_2. First, an automatic mapping algorithm A produces a set of candidate mappings between O_1 and O_2. Each candidate mapping m represents a potential correspondence between a concept in O_1 and a concept in O_2. The concepts can be classes, properties, or axioms in the ontologies. Correspondences are typically an equivalence or a similarity relation ($=$), but can be a subsumption relation ($<=$, $>=$), or any other (domain-specific) relation. In the current implementation of CrowdMap, we consider only $=$, $<=$, and $>=$. The algorithm A may also produce a confidence measure $conf$. If A does not produce confidence measures, then we assume that $conf = 1$ for all mappings returned by A.

We generate microtasks as follows.

- There is a microtask to verify each candidate mapping m. Tasks can either ask workers either to validate a given mapping relationship between the source and target (such as similarity), or to choose between different types of relationships between the source and the target (such as subsumption, similarity, or meronymy).

- If the algorithm A produces only equivalence (similarity) mappings, then CrowdMap requests 3 workers to verify the same mapping.

- If the algorithm A produces equivalence and subsumption mappings, then CrowdMap asks for up to 7 workers to complete the task of selecting a relationship between the source and target, until at least two of them agree on a choice of relationship between the two terms.

- The final set of mappings is the set of mappings M_c where at least 2 workers agreed on the type of the mapping.

The number of workers that we assign for each microtask is a configuration parameter. The values that we used in the current version of CrowdMap follow common practice in using microtask platforms for similar types of tasks. We assume that a higher number of answers are required to validate the second type of task (asking for equivalence and subsumption), which is significantly more complex from an alignment point of view and has more options for workers to choose from. Our pilot studies helped us determine others, such as the choice of words and methods to avoid spam (Section 3.5).

3.3.3 The CrowdMap Architecture

Figure 3.1 shows the CrowdMap architecture. The dashed line separates the modules that prepare and publish microtasks from the modules that process the responses of the crowd. CrowdMap executes the former set of modules first (see the specific order in the numbers). Once CrowdMap creates the microtasks in CrowdFlower and they are published to the actual labor platforms such as MTurk, the crowd interacts with the MTurk interface and provides responses to the microtasks. When CrowdFlower receives the full set of answers for these microtasks, CrowdMap executes the second set of modules and calculates the resulting alignment.

Mappings Generator The current CrowdMap prototype focuses on pairs of classes as elements to be compared through crowdsourced alignment. We do not yet support mappings between properties, but many of the main findings of our experiments are likely to apply to these types of ontological primitives as well. The Mappings Generator processes the alignment from an automatic tool or uses one of its

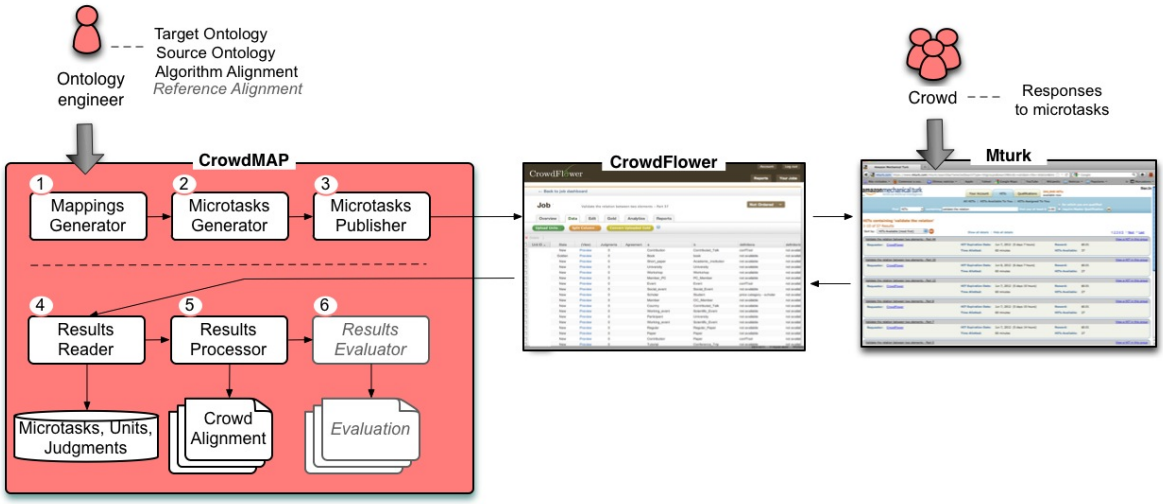

Figure 3.1: CrowdMap architecture. CrowdMap generates microtasks using a set of pairs of ontological elements and the relationships between them, publishes the microtasks to CrowdFlower, retrieves the answers of the crowd, and compiles the final alignment results by deciding which of these answers are valid.

benchmark-generation mechanisms to generate a set of mappings to test. Section 3.4 discusses the different sets of candidate mappings that we generated for the experiments.

Microtasks Generator This module generates the microtasks associated with the pairs of classes computed by the Mappings Generator. We can further parameterize the process by configuring such aspects as interface and layout, number of answers for each alignment question, number of questions within one microtask, and restrictions on the workforce (e.g., a certain level of performance achieved so far, geo-location, language skills). The result is the actual interface that the workers will use in order to submit their answers.

Microtasks Publisher The publisher module posts the microtasks to the crowdsourcing platform. In the current implementation we support the publication to CrowdFlower using the API that it provides. The publisher module creates the corresponding microtasks on CrowdFlower, uploads the data about the normal and the golden units, and publishes the microtasks on MTurk.

Results Reader Once the microtasks are completed, CrowdFlower calculates an aggregated response for each pair of terms to align, as well as the confidence of such aggregated responses. The confidence combines the accuracy that workers obtained in the microtask with the agreement of the responses for the alignment question at hand. Access to this information is provided through the CrowdFlower API.

Results Processor This module generates a file with the crowd alignment, serialized in the Alignment API format (David et al., 2011). The usage of this standard format facilitates the comparison between different approaches (crowdsourced vs. automatic, reference data vs. manually or automatically generated), as well as the reuse of the results in new scenarios involving both human-oriented and algorithmic processing.

Results Evaluator The evaluator module relies on the Alignment API to assess the crowd alignment. Via the API we access information about specific alignments (the ones computed by the crowd, and reference alignments) and compute precision and recall values.

The functionality offered by CrowdMap could be easily integrated into existing environments for

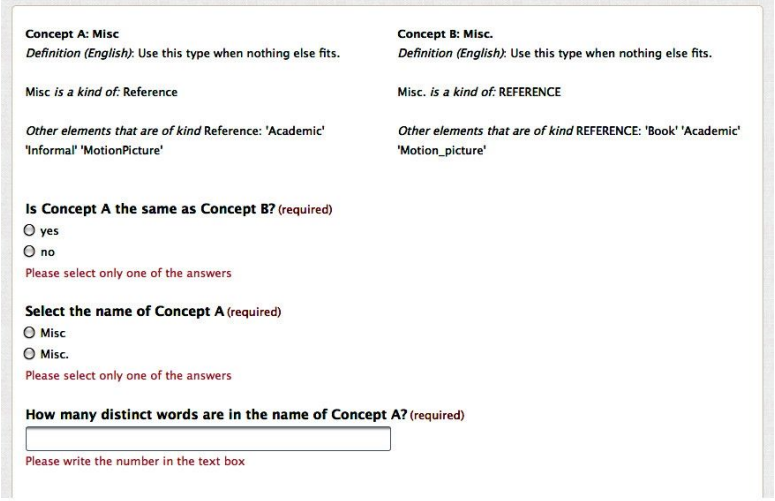

Figure 3.2: User interface of a validation microtask. CrowdMap shows the worker two elements to be aligned and asks whether they are related to each other with a particular relationship.

ontology alignment, such as the PROMPT Protégé plug-in (Noy and Musen, 2003) or even used to complement tools that perform data interlinking, such as Silk (Volz et al., 2009) and Google Refine with curated information about schema-level alignments.

3.3.4 Microtask User Interface Design

In CrowdFlower, the user interface that a worker sees has three main parts: (i) the title and instructions explaining the purpose of the microtask; (ii) the problem statement, which in our case is the information about the elements (e.g.,classes) to be compared; and (iii) the form that workers must fill out to submit their responses. CrowdMap defines two types of microtasks for which we generate different interfaces: (i) validation microtasks and (ii) identification microtasks. A validation microtask presents workers with a complete mapping (e.g., two classes and the relationship that connects them) and asks them to specify whether they agree with the relationship that they see. An identification microtask asks for workers to identify a particular relationship between the source and the target classes. Figure 3.2 shows an example of a validation microtask. The first part is the problem statement; the second part is the form. The microtask includes all contextual information available for both classes (labels, definitions, superclass, siblings, subclasses and instances). The first element in the form asks the user whether or not the concepts are similar. The form also includes two more elements as verification questions that help in filtering spam, similarly to the approach by Kittur and colleagues (Kittur et al., 2008). We use a different input form for identification microtasks. Figure 3.3 shows the first field of three sample questions within an identification microtask. CrowdMap can create identification microtasks showing either a complete version of the form (relationships $=$, $<=$, $>=$, none), or a short version ($=$, not $=$). Anti-spam mechanisms are the same as for validation microtasks, illustrated in Figure 3.2.

In order to reduce response bias, CrowdMap creates only half of the HITs using the interface in Figures 3.2 and 3.3. In the other half, CrowdMap presents the possible answers in the opposite order, and focus the verification question on the other class in the pair to be matched. This technique, which we apply independently from the type of microtask, makes the evaluation of workers stricter, allowing us to identify and block spam more efficiently. The verification questions that we used to identify and

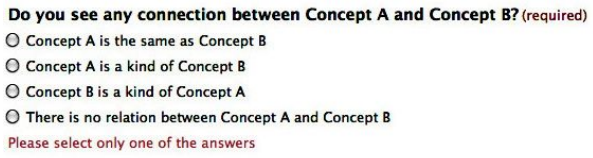

Figure 3.3: User interface of an identification microtask where CrowdMap shows the worker two elements to be aligned and asks to identify the relationship between them. The relationship in this case can be that both are the same, one is more specific than the other, or the two are not the same

avoid spam play a special role in these checkpoint-like questions; the response of a worker to a golden unit is evaluated positively only if all three fields of the input form have a correct response.

3.4 Evaluation

In order to perform our analysis, we conducted several studies to test both the feasibility of overall approach and specific characteristics of the design of crowdsourced ontology alignment that improve its effectiveness. We used the ontologies and the reference alignments from the Ontology Alignment Evaluation Initiative (OAEI) as golden standard to assess the accuracy of the crowd-computed results.

3.4.1 Ontologies and Alignment Data

We have conducted three sets of experiments in order to address the research questions from Section 3.1 (Table 3.1).

In our first experiment, CartP, candidate mappings included all possible pairs of mappings between two input ontologies (a Cartesian product of the sets of classes). While such an approach does not scale in practice, it provides the baseline on the best possible performance (recall in particular) of crowdsourced alignment. The OAEI ontologies that we use for the CartP experiment are two ontologies that cover the BibTex data, one from MIT and one from INRIA (ontologies 301 and 304 from the OAEI set). For each pair of classes we provide the user with contextual information that is relevant to the corresponding elements and compare the results against the reference alignments provided by the OAEI.

The second type of microtasks, which we call Imp, uses only those class pairs that were created by a given ontology alignment tool as a set of candidate mappings. This experiment simulates a typical CrowdMap workflow (Figure 3.1). We used the output of the AROMA tool as our input alignment. AROMA is one of the algorithms from OAEI that presented a good performance in 2011. Again, we ran the experiment using ontologies 301 to 304 just as in the CartP and included full context-specific descriptions of the two elements to be matched. Note that we obtained the results for the Imp setup by using the CartP data since we already had the judgements for all the pairs of terms from the two ontologies that we used in both experiments.

The third set of microtasks, which we call 100R50P, includes several ontology pairs and allows us to compare the CrowdMap performance in different settings. The sets of candidate mappings in the 100R50P experiments simulate input originating from a tool with 100% recall and 50% precision. We create the set of class pairs where 50% of the mappings are correct and 50% are incorrect. We take the correct mappings from ontology alignment reference data. Incorrect mappings consist of false

negatives (generated by an algorithm), as well as randomly generated mappings. If there is no algorithm to generate candidate alignments, we generate all the incorrect mappings by selecting pairs of classes randomly.

We use the *Conference ontologies* from the OAEI set. The ontologies in this set represent knowledge about conferences and were produced by different organizations. Some of the selected ontologies are based on actual tools for conferences (Cmt and ConfOf), and others are based on either personal experiences (Ekaw) or Web pages of conferences (Sigkdd). We took a pair of ontologies from this set, choosing the Argmaker algorithm results as the alignments performed by the automatic tool.

The ontologies in the OAEI *Oriented matching* set cover the domain of academia and the reference alignment includes complex relationships, such as broader than and narrower than. We took the same pair from this set that we used in the CartP experiment (301 to 304).

Table 3.1 summarizes the three experiments.

	CartP	**Imp**	**100R50P**
Ontologies	301-304	301-304	Edas-Iasted, Ekaw-Iasted, Cmt-Ekaw, ConfOf-Ekaw
Input alignment	Cartesian product	Output of the AROMA algorithm	50% correct mappings (all mappings from the reference alignment), 50% incorrect mappings, output of AgrMaker
Research question	R1	R2	R2, R3

Table 3.1: Summary of the experiments

3.4.2 CrowdFlower and MTurk Setup

Both CrowdFlower and MTurk allow requesters to configure their microtask projects according to a number of different parameters. In our experiments, we clustered 7 different alignment questions (or units in CrowdFlower parlance) into one HIT. This step facilitates worker assignment and resource optimization (see Section 3.3.1). One of these questions was a golden unit (see Section 3.3) where we knew the answer in advance. We could use it to assess the performance of workers, to deal with spammers, and to validate the final results. For each experiment we selected golden units from a set of 50. Each HIT includes two verification questions, which apply to both golden and real units, as a means to reduce spam (see Section 3.3.4).

Redundant answers to the same question are a useful way to evaluate the feasibility of the overall approach—can users actually agree on the answer?— and to (automatically) identify correct answers. We requested 3 workers for those questions that asked them whether a given correspondence holds or not. These values are based on best practices in crowdsourcing literature (Ipeirotis et al., 2010).

It is common for microtask platforms to organize HITs in batches. In our case, each batch contained at most 50 HITs, each with 7 mapping units. This value is an empirical one used in similar experiments on MTurk (Kittur et al., 2008), which balances resource pooling and the time required to complete a full batch. Several workers verified each alignment, not only to receive the minimal number of answers required for majority voting, but also because we wanted to change the order of the allowed answer choices to avoid spammers. We calculated the number of golden units as the number of HITs in each group, and adjusted the number of mappings to show in each set of alignment questions, in cases where it was needed by the CrowdFlower internal restrictions. CrowdFlower requires that a worker answers 4 golden units correctly before she becomes a trusted workers. We reduced this number to 2 since we

	CartP 301-304	100R50P Edas-Iasted	100R50P Ekaw-Iasted	100R50P Cmt-Ekaw	100R50P ConfOf-Ekaw	Imp 301-304
Precision	0.53	0.8	1.0	1.0	0.93	0.73
Recall	1.0	0.42	0.8	0.75	0.65	1.0

Table 3.2: Precision and recall for the crowdsourcing results

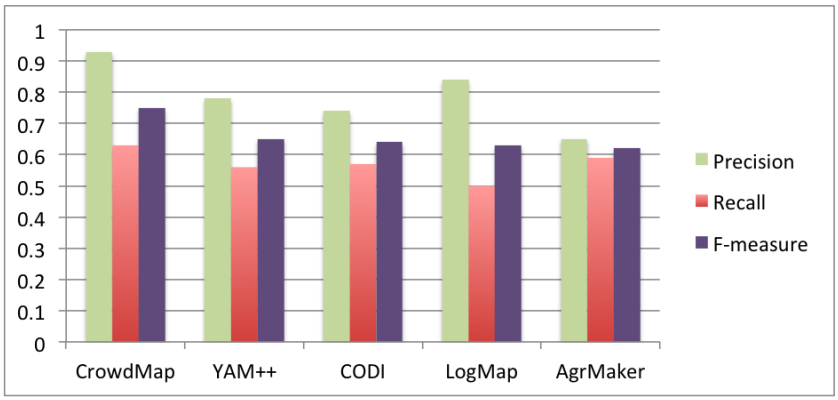

Figure 3.4: The average precision, recall, and F-measure of CrowdMap and the top performers on the conference set for OAEI 2011 (http://oaei.ontologymatching.org/2011/results/conference/index.html)

observed that workers were submitting fewer than 4 correct answers to golden units, even with correct mapping results.

For most experiments we paid $0.01 for each HIT; for the CartP scenario we raised the reward to $0.04 to compensate for the larger scale of the experiment and to study the trade-offs between time to completion and costs. CrowdFlower publishes jobs on the platform for 7 days by default and the deadline is extended until jobs are completed. For most of the experiments we needed between 7 and 10 days, which is possibly also a consequence of the fact that we published several similar jobs within a relatively short period of time. The higher-rewarded experiments required less than a day to finalize, which was significantly faster than other trials we ran on the same data and $0.01 per HIT.

3.4.3 Results

Table 3.2 shows the precision and recall in our experiments. We use the PRecEvaluator available in the Alignment API to calculate these values.

The results show very high precision for the conference alignments. Figure 3.4 compares the performance of CrowdMap on the conference set with the 4 top performers in OAEI 2011. The chart shows the average precision, recall, and F-measure. Note that CrowdMap significantly outperforms the other algorithms, with the F-measure of 0.77. It is important to note, however, that for the conference set, CrowdMap does not start with a cartesian product of all possible pairs. It needs to filter only a set of mappings that have 50% correct mappings and 50% wrong mappings. However, the crowd improved the precision considerably from that 50%.

For the CartP alignment, the workers have found all the mappings from the reference alignment,

achieving a remarkable 1.0 recall. The precision, however, has suffered. We address this issue in our discussion in Section 3.5.

3.5 Analysis and Lessons Learned

The results of the experiments lead to the following conclusions

R1 From the result achieved in the CartP experiment we can conclude that our approach is feasible. Given the full set of potential correspondences between pairs of classes, the crowd was able to provide meaningful answers that could be used in the alignment process.

R2 If we compare the results of 100R50P with the performance of the AgrMaker algorithm which we used as a baseline on the conference alignments (precision: 0.65 and recall: 0.59), we notice that CrowdMap can improve both the precision and the recall of the original algorithm. This finding is supported by the outcomes of the Imp experiment, by comparison with performance of the AROMA tool on the benchmark alignments (precision: 0.93 and recall: 0.53).

R3 Workers were capable of submitting correct responses with both validation and identification microtasks.

One unexpected observation from Table 3.2 is the effect of the number of mappings that we present on precision. The precision is very high for all the tasks in the 100R50P set, where we showed only a limited number of pairs of mappings. In the CartP experiments, the workers had access to the cartesian products of all the class pairs, and the precision dropped significantly. Because we used the CartP results to simulate the Imp experiment, the precision there suffered as well. Our hypothesis for this low precision is that the large task might attract more spammers or more workers just try to get through the task quickly. However, in future work, we plan to design experiments to test this hypothesis.

However, if we look at results from the different experiments together, we can see a potential for a two-step process that might be very efficient. The workers can achieve perfect (or close to perfect) recall when given a large set of candidate pairs, many of which are not mappings. They achieve high precision on a set that has fewer wrong mappings but all correct ones. Thus, we can use a setting such as CartP (extremely low precision, perfect recall) to get a set that is close to 100R50P (50% precision, 100% recall). Indeed, CartP produced a mapping set that was extremely close to an 100R50P set. This approach would create a two-step CrowdMap algorithm: first stage uses CartP (or its approximation, by taking all the very low confidence mappings from an automatic tool). Then we can use the results of this first stage as an input to another run of CrowdMap which will improve the precision. Note that this approach is similar to the Find-Fix-Verify crowd programming pattern in Soylent (Bernstein et al., 2010).

We carried out the experiments over a period of five weeks, whereas half of this time was dedicated to the tuning of the configuration parameters of the crowdsourcing platform and the testing of different variants of the interfaces (see Section 3.3.4). In its current, optimized version, we estimate that CrowdMap could produce accurate alignments between pairs of ontologies within a relatively short period of time (around one week for several hundreds of HITs and corresponding alignments). The total costs of the experiments were around $50, which is not comparable to alternative approaches oriented at knowledge engineers or domain experts, with or without the involvement of automatic algorithms.

Before running the experiments that we reported, we tested the prototype with small pilots. The pilots allowed us to fine-tune the user interface and to develop methods to minimize spam. When we

initially did not use golden units or verification questions, we received a huge amount of spam. While we collected the required responses in a few hours, most of them appeared to be very low quality ones. Over several iterations, each of which reduced the number of spam, we came to the following strategies. First, we use golden units to block invalid answers. Second, we use verification questions that force the user to type a name of the concept. Finally, CrowdFlower allows requester to exclude specific countries that have workers who tend produce the majority of spam answers. Including developing countries such as India was another strategy that helped reduce spam significantly.

The wording and structure in the user interface also influenced the results. We experimented with different types of verification questions and phrasings thereof. We wanted to define additional questions that were trivial to answer, yet, required the user to process cognitively the information on the form. We also needed verification questions that would get different answers from one pair of terms to the next, so that workers could not cut and paste. In the experiments that we report here, we used both the names of the classes to be compared, as well as other features such as the number of words in the class names as basis for such verification questions. For one type of verification question asking for the name of one the classes to be matched, we eventually decided in favor of a radio button rather than a free-text field, as in the latter case many workers simply typed in the default name 'Concept A' mentioned in the question. References to the "first" or "second" class in the matching pair also turned out to confuse users. In the case of a second verification question, which asks about the number of distinct words displayed, a simple validator encouraged workers using positive integers (e.g., "1") instead of text (e.g. "one"), and thus avoiding correct responses to be evaluated negatively. Changing the wording of equivalence-alignment questions from "Concept A is similar to Concept B" to "Concept A is the same as Concept B" lead to a better understanding of the task by the workers and to better results. Finally, we verified how important ontology documentation is, since CrowdMap relies on the quality of labels and definitions.

Another observation that we made is related to the number of related microtasks (or groups of questions) published at the same time; in this case the time to completion increased, probably due to the fact that the same workers typically take the opportunity to solve a series of similar tasks. The results that we have obtained largely depend on the dataset used for the evaluation. It is worthwhile mentioning that, there have been cases in which the crowd identified mappings that were correct in our opinion (such as $Person - Person$), but were not present in the reference alignment. This means that these mappings did not count for the recall and precision values. We also analyzed the mappings that the crowd missed from the reference alignment, and we must say that there were cases that were not clear for us either. For example, mappings such as $WelcomeTalk - Welcome\_address$, or $SocialEvent - Social\_program$, or $Attendee - Delegate$ (from test $Edas - Iasted$) are ambiguous.

Most work on using crowdsourcing for computational tasks rely on MTurk as a platform. Our experiences with CrowdFlower showed that this platform represents a real alternative to directly accessing the MTurk crowd, in particular due to the additional features they offer with respect to quality assurance. However, it is worthwhile mentioning that while it is possible to use MTurk via CrowdFlower, the latter does not support the full range of services of the former; for instance, it is not possible to update the number of answers required for a question during the execution of a task.

3.6 Conclusions and Future Work

This paper makes several contributions to the state of the art in ontology alignment. First, we present a workflow model for crowdsourcing ontology mappings and describe the implemented solution that uses

CrowdFlower. Second, we perform a feasibility study for the use of crowdsourcing to perform ontology mapping. Third, we provide an analysis of the characteristics of crowdsourced ontology mappings for different ontologies, mapping relationships, and settings. Our first prototype of CrowdMap has proven that the crowdsourcing approach to ontology alignment is feasible, and can augment automatic tools in a cost-efficient, fast, and scalable manner.

Future work will focus on executing new experiments to analyze further research questions. For example, we would like to discover which contextual aspects are the most useful to improve accuracy, and whether we could use agreement among workers to determine the certainty of mappings. We expect to create a set of instances for each ontology used in the experiments, so that workers can see up to 5 instances as part as the context of the elements to be aligned. We will perform more experiments to test whether accuracy is reduced in cases where the domain of the ontologies requires specific knowledge (e.g., biomedical ontologies). Finally, after completing the extensive set of experiments, we believe that we can improve the worker performance by fine-tuning the question wording even better (e.g., substituting the class names directly into the options for selection in the mapping questions).

We plan an extension of the implemented prototype of CrowdMap to enable crowdsourced mappings between ontology properties and axioms. With respect to the actual workflow, we will look into more sophisticated means to combine the results of human and algorithmic computations, by following, for instance, a Bayes analysis approach (cf. (Demartini et al., 2012)). Along the same lines, we also intend to apply filtering techniques to optimize the number of questions that are issued to the crowd to improve scalability and costs. Such filtering is an essential pre-requisite for the application of CrowdMap to related fields such as data interlinking, which has orders or magnitude more data and possible a larger degree of noisy data than the scenario that we studied in this paper.

Acknowledgements

We would like to thank the self-service team of CrowdFlower, for their technical support on the CrowdFlower API.

Chapter 4

Intrinsic Measures for Link Quality Assessment: Information Gain Enabled by Links

Abstract: The current Web of Data contains a large amount of interlinked data. However, there is still a limited understanding about the quality of the links connecting entities of different and distributed datasets. Our goal is to provide a collection of indicators that help assess existing interlinking. In this paper, we present a framework for the intrinsic evaluation of RDF links, based on core principles of Web data integration and foundations of Information Retrieval. We measure the extent to which links facilitate the discovery of an extended description of entities, and the discovery of other entities in other datasets. We also measure the use of different vocabularies. We evaluated our link assessment measures using links extracted from a set of datasets from the Linked Data Crawl 2014, in terms of measure validity, feasibility according to real-world data and usefulness for users in the role of data publishers.

4.1 Introduction

Linked Data principles encourage data publishers to connect the resources in their datasets to other resources "so that more things can be discovered"[1]. With the increasing number of available datasets and links between them (Max Schmachtenberg and Cyganiak, 2014; Schmachtenberg et al., 2014), it becomes highly important to observe the extent to which existing links have desirable properties, as we need to ensure high quality to encourage the discovery and usage of Linked Data. Not only of the datasets as silos but of the links as well

Links should (i) follow the recommendations that apply to high quality data (Zaveri et al., 2016) (i. e. links should be accessible, syntactically valid, and semantically accurate), and (ii) links should enable the discovery of "more things", facilitating new insights from the data. Established data-driven quality assurance methodologies (Pandian, 2003; Zaveri et al., 2016; Rula and Zaveri, 2014) suggest that the key steps for improving the status quo are: the definition of measures, the analysis of measurements and the subsequent monitoring of updates. So, to be able to analyse the quality of links, we need measures that help us assess all relevant quality aspects, including (i) and (ii).

Previous empirical studies on the adoption of Linked Data principles (Schmachtenberg et al., 2014; Hogan et al., 2012) report on the number of outgoing and incoming links of datasets, and the most frequently used predicates in RDF links. Recently, Hu et al. (2015) studied degree distributions, as well as missing links in Bio2RDF based on symmetry and transitivity. Neto et al. (2016) focused on the analysis of dead links in schema and entity link triples published in the Web of Data. While these studies, together with the findings provided by smaller evaluations of other link assessment methods focusing on (i) (e. g. Guéret et al. (2012) and Albertoni

[1] Berners-Lee, T. Linked Data Principles `http://www.w3.org/DesignIssues/LinkedData.html`

and Pérez (2013) provide a characterization of existing links), they do not allow for assessing how many new things might be made discoverable thanks to the links (ii).

In this paper, we provide a framework for link analysis that takes into account principles of data integration in the Web of Data, addressing (ii). We suggest measures that focus on data quality dimensions inherent in the data, while extrinsic assessment would take into account the needs a user has in his specific context (cf. (Zaveri et al., 2016)) More specifically, our measures examine the effect that links have on entities (and consequently on datasets). We measure the extent to which links facilitate the discovery of an extended description of entities, and the discovery of other entities in other datasets. We also measure if they add different vocabularies (cf. Section 4.4.2) to the description of entities. Our measures are grounded on foundations of the field of Information Retrieval, as we acknowledge redundancy when we measure the gain in description, connectivity and number of used vocabularies. More precisely, the contributions of this paper are:

1. We identify a set of principles for data interlinking in the Web of Data (Section 4.3).
2. We define a set of measures to analyse available links in terms of these principles (Section 4.4).
3. We demonstrate the feasibility of the proposed framework with the implementation of the measures and carry out an empirical analysis of links extracted from the Linked Open Data Crawl (Schmachtenberg et al., 2014) (Section 4.5).

4.2 Preliminaries

We introduce in this section the terminology and notation. In order to increase the readability of the following text, we re-introduce some of the definitions presented in Chapter 2 that are relevant to the content in the present chapter.

Definition 4 ***RDF Quadruple***: Given $\mathcal{U}$, a finite set of HTTP URIs, representing resources, $\mathcal{L}$ a finite set of literal values, and a finite set of blank nodes $\mathcal{B}$ where $U \cap \mathcal{L} = \mathcal{U} \cap \mathcal{B} = \mathcal{L} \cap \mathcal{B} = \emptyset$, a quadruple (s, p, o, c) is any element of the data space $Q = (\mathcal{U} \cup \mathcal{B}) \times \mathcal{U} \times (\mathcal{U} \cup \mathcal{L} \cup \mathcal{B}) \times \mathcal{U}$. s, p, o is a triple statement describing s, while c is the context (denoted by a URI) in which the triple is defined.

Definition 5 ***RDF Dataset with Context***: An RDF dataset D_c is a set of quadruples grouped by some context c $D_c \subseteq \{(s, p, o, c) \in Q\}$, where Q is the set of all quadruples.

Definition 6 ***Home***: Given C the set of all contexts, and an entity (either a blank node or URI), $home : \mathcal{B} \cup \mathcal{U} \mapsto C$ is the function that maps the entity to the context c where the entity is defined. Note that when x is a vocabulary term (e. g. a class or a property), the c returned by $home(x)$ is the identifier of the vocabulary where the term x was defined.

The $home$ function is customizable. For example, it can be defined to match the notion of datasets in the Linked Open Data literature (Schmachtenberg et al., 2014), or it can be defined to match the graphs in datasets—the graphs in the SPARQL and N-Quads specifications. In this paper, we stick to the LOD cloud diagram[2] and analyse links on a dataset basis.

For representing the relation between entities of different datasets, we define:

Definition 7 ***Link***: A link of D_c is a quadruple $(s, p, o, c) \in D_c$ such that $s \in \mathcal{U}, o \in \mathcal{U}, home(s) = c, home(o) \neq c$.

Definition 8 ***Interlinking***: The interlinking I_c of a dataset D_c is the set of all links going out from D_c to any other dataset: $I_c = \{(s, p, o, c) \in D_c | \ home(s) = c, home(o) \neq c\}$.

[2] `http://lod-cloud.net`

To formally define our measures, we use a relational algebra-like notation. For this purpose we define selection σ, projection π and join $\bowtie$ as follows:

Definition 9 ***Selection****: Given $X \subseteq D_c$, a selection $\sigma_h(X)$ is the quadruples from X that satisfy a selection predicate h: $\sigma_h(X) = \{(s,p,o,c)|(s,p,o,c) \in X \wedge h(s,p,o,c) = true\}$*

Example 1 *: We can select the quadruples of the dataset D_c that are `owl:sameAs` links by $\sigma_{p=owl:sameAs}(D_c) = \{(s,p,o,c)|(s,p,o,c) \in D_c, p = owl:sameAs\}$*

Definition 10 ***Projection****: Given $X \subseteq D_c$, and Y a subset of the elements in the quadruples in X, a projection $\pi_Y(X)$ on attributes Y is the subset of X including the elements Y: $\pi_Y(X) = \{(s,p,o,c)[Y]|(s,p,o,c) \in X\}$*

Example 2 *: We can obtain the projection of all the entities appearing in the predicate and object positions of the quadruples of the dataset D_c by $\pi_{p,o}(D_c) = \{(p,o)|(s,p,o,c) \in D_c\}$*

Definition 11 ***EquiJoin****: Given $X_1 \subseteq D_1$ and $X_2 \subseteq D_2$, the Equi join of the two sets is the set of elements such that: $X_1 \bowtie_{X_1.o\theta X_2.s} X_2 = \{(X_1.s, X_1.p, X_1.o, X_2.s, X_2.p, X_2.o, c) \mid X_1.o = X2.s\}$*

Example 3 *: In Table 4.1 case (I), the equijoin of the two quadruples on the name and the link is the 7-tuple "d1:nn owl:sameAs d2:nn d2:nn rdfs:label "Natasha" d1 .".*

Now, we may re-state **our task** at hand as follows: Given a dataset D_c containing the interlinking I_c, our task is to compare D_c and $D_c \backslash I_c$ and analyse the value that I_c gives to the data in terms of the principles for data interlinking in the Web of Data described in the following section.

4.3 Principles for Data Interlinking in the Web of Data

The main reason to connect datasets is to enable their joint search, browsing or querying. As in any information system, when a user queries Linked Data it is important that she: (**n1**) finds all entities she is interested in (recall); (**n2**) finds only entities she is interested in (precision); (**n3**) is able to understand the relationship between entities in the Web; (**n4**) finds answers to all her questions no matter how heterogeneous in syntax, structure and semantics the questions are.

The existence of high quality links between entities can contribute to a better fulfilment of the aforementioned needs (n1-n4). In order to understand the way links can help, let us consider various interlinking examples (from (I) to (VII)) shown in Table 4.1. We analyse each of the examples, and derive from them desired properties for links (i. e. principles P1-P3).

Entity Description In case (I) we see two entities linked via an `owl:sameAs` link. The two connected entities have different names, but represent the same person (Natalya F. Noy, also known as Natasha Noy informally). The source dataset contains the publications that Natasha wrote when she worked at Stanford, and the target dataset contains publications she has written while working at Google Inc. If we search for the publications written by Natasha and only consider the source dataset, we exclusively see her Stanford publications. If we consider the link connecting the two entities referring to Natasha, we are able to also find her Google publications, giving us higher recall (**n1**).

In case (II) the two entities are also connected via an `owl:sameAs`. The target dataset contains data about conferences and program committees, while the source dataset does not contain this kind of data. If we look for persons who have been chairs of scientific events, and we only take into account the source dataset, we are not able to find any person because we lack the information about the chairs of the events. In an Information Retrieval scenario, we would use query relaxation techniques (e. g. (Fokou et al., 2016; Mottin et al., 2014)), and the search query would be reformulated as a search for persons. The result would include the entities for Natasha Noy and Mark Smith (who is a student assistant and was never a chair). Conversely, if we consider the link, we

	Source dataset	Target dataset(s)
	Entity Description	
(I)	d1:nn foaf:name "Natasha Noy" . d1:nn dbo:affiliation d1:stanford . d1:nn swrc:publication d1:p2012-1 . d1:nn owl:sameAs d2:nn .	d2:nn foaf:name "Natalya F. Noy"" . d2:nn dbo:affiliation d1:googleinc . d2:nn swrc:publication d2:p2015-1 .
(II)	d1:nn foaf:name "Natasha Noy" . d1:nn owl:sameAs d2:nn . d2:ms foaf:name "Mark Smith"	d2:nn foaf:name "Natalya F. Noy" . d2:nn cito:likes d2:sfo . d2:nn swc:holdsRole swc:Chair
(III)	d1:nn foaf:name "Natasha Noy" . d1:nn owl:sameAs d2:nn .	d2:nn foaf:name "Natasha Noy" .
	Entity Connectivity	
(IV)	d1:nn foaf:name "Natasha Noy" . d1:nn dbo:affiliation dbr:Stanford_University . d1:nn owl:sameAs d2:p1 . d1:nn owl:sameAs d3:p5 . d1:nn owl:sameAs d4:p1 .	d2:p1 foaf:name "Natasha Noy" . d2:p1 dbo:affiliation dbr:Stanford_University . d3:p5 foaf:name "Natasha Noy" . d3:p5 dbo:affiliation dbr:Stanford_University . d4:p1 dbo:affiliation dbr:Stanford_University .
(V)	d1:nn foaf:name "Natasha Noy" . d1:nn dbo:affiliation dbr:Stanford_University . d1:nn owl:sameAs d2:p1 .	d2:p1 foaf:name "Natasha Noy" . d2:p1 dbo:affiliation dbr:Stanford_University . d3:p5 foaf:name "Natasha Noy" . d3:p5 dbo:affiliation dbr:Stanford_University .
	Vocabularies Involved in the Description	
(VI)	d1:nn foaf:name "Natasha Noy" . d2:nn rdf:type foaf:Person . d1:nn owl:sameAs d2:nn .	d2:nn sioc:creator_of d2:post2 . d1:nn rdf:type proton:Human . d2:nn vivo:teachingOverview "Natasha Noy was a tutor in the SSSW08 summer school" .
(VII)	d1:nn foaf:name "Natasha Noy" . d1:nn owl:sameAs d2:nn .	d2:nn foaf:name "Natasha Noy" . d2:nn foaf:currentProject d2:bioportal . d2:nn foaf:pastProject d2:protege .

Table 4.1: Examples of different interlinking cases.

have relevant information for the query and only Natasha is retrieved in the results. Therefore, in this case the link enables us to have higher precision (**n2**).

Observation: These two cases, have something in common: the links (s, p, o, c) extend the description of entities s. The description of an entity is the set of quadruples with s as subject, and literals, URIs and blank nodes as objects (cf. Section 4.4.2). When the linked datasets provide redundant information, links do not help in recall, nor in precision. Example (III) is a clear example of a scenario where we have redundant information and the description is not extended. Therefore, we formulate the first principle as:

P1: Try to create an interlinking that extends the description of entities of the source dataset.

Entity Connectivity Case (IV) connects the entity referring to Natasha in d1 to the corresponding entities representing Natasha in datasets d2, d3 and d4. While these links do not extend description of the entity in d1 (i. e. they do not follow the Principle P1), they help in understanding the relationship between the entities in the Web of Data (**n3**). This understanding is necessary when for example, a change in the affiliation of Natasha is materialised in d1 to update her affiliation. The descriptions in d2, d3 and d4 could be subsequently changed, in order to keep the data up-to-date.

Observation: In (IV), we can see the importance of creating multiple links from the same entity to different external entities and datasets, increasing its connectivity (cf. Section 4.4.2). In Case (V), which is similar to case (IV) but without the links to d3 and d4, we see that if the links from d1:nn to the entities in d3 and d4 do not exist (as in case (V)), it is harder to reach the entities in other datasets that would need to be updated. This is similar in cases where the links are created to group entities, or to enable the browsing of different types of entities. We formulate the second principle as:

P2: Try to create an interlinking that increases the number of entities and datasets that source entities are connected to.

Heterogeneity of Descriptions Case (VI) shows an example where the entity representing Natasha is connected via an `owl:sameAs` link to its corresponding entity in d2. The entity in d2 is described with vocabularies that are different from d1's vocabularies. In contrast, in case (VII) the entity in d2 contains a description that adds new information to the description of d1 (satisfies P1) but uses the same vocabulary as in d1 (i. e. FOAF).

Observation: in (VI), links help in answering a wider range of queries that might be formulated in different application contexts (**n4**). Using different vocabularies we are able to use and analyse entities from multiple perspectives. Hence, the third principle is:

P3: Try to create an interlinking that makes the source entities have a description with a higher number of vocabularies in their description.

These principles are not independent from each other. Principles P2 (entity connectivity) and P3 (vocabularies) are specializations of P1 (entity description). For some types of links (non-identity links), creating links to new entities in new datasets (P2), means that the description of the source entity is extended (P1). However, that does not necessarily happen the other way round. Analogously, if one uses further vocabularies in the links between entities (P3), the description of the source entity will be extended (P1). We do not claim that these principles are complete, and they may be extended.

4.4 Intrinsic Measures for Assessing the Quality of Links

The measures that we define do not provide an absolute assessment of the quality of links. That is, a particular measurement is not good or bad. Instead, we provide measures for a comparative assessment: we acknowledge that one interlinking is better than another in some dimension that we observe with regard to the principles in the previous section. It is up to the person or application inspecting the measurements to interpret its meaning, and make a decision based on it (e. g. a data publisher willing to improve her interlinking and using our measurements as a guide to decide where to start from).

We distinguish between descriptive statistics that give an overview of the size and the elements in I_c (see Section 4.4.1), and measures that assess the way the links in the interlinking I_c of the dataset D_c follow the aforementioned principles (see Section 4.4.2).

4.4.1 Basic Descriptive Statistics

In order to describe basic properties of the interlinking of a dataset, we use basic statistics proposed by related work (e. g. Void Vocabulary[3] and LOD Stats[4]), to compute the volume of the interlinking ($|I_c|$), and the distribution of linksets ($\{(x, |\sigma_{p=x}(I_c)|)\}$).

[3] `https://www.w3.org/TR/void/`
[4] `http://stats.lod2.eu/links`

4.4.2 Principles-based Measures

Since we would like to study the effect that links have on the entities of the source dataset, our measures analyse links grouped by source entities. Note that in our analysis we focus on entities $e \in D_c$ such that $\nexists\ (e, rdf : type, rdfs : Class) \in D_c$. So, we look at the interlinking of individuals and not at vocabulary terms.

4.4.2.1 Two views of the quadruples about entities

For each entity e, we distinguish two views of the set of quadruples that state something about e: the description view and the connectivity view of an entity.

4.4.2.2 Description view

This view focuses on all the quadruples in X describing the entity e.

We define the description of an entity e in $X \subseteq D_c$ as the projection that selects the predicates and objects from the set of quadruples of X about e, and entities defined to be identical to e (usually defined via the predicates `owl:sameAs` or `skos:exactMatch`).

$$desc(e, X) = \pi_{(p,o)}(\sigma_{s=e}(X)) \cup \pi_{(Q.p,Q.o)}(\sigma_{X.p=identity}((X \bowtie_{X.o=Q.s} Q))) \tag{4.1}$$

In order to have a more detailed view of the description, we differentiate between the entity's classification (i. e. the quadruples referring to the `rdf:type` of the entity):

$$classif(e, X) = \sigma_{p=\text{“}rdf:type''}(desc(e, X)) \tag{4.2}$$

and the rest of the description:

$$descm(e, X) = desc(e, X) \backslash classif(e, X) \tag{4.3}$$

Example 4 *In Table 4.1(VI), classif(d1:nn,D_1'= { (rdf:type, foaf:Person), (rdf:type, proton:Human)} and descm(d1:nn,D_1') = {(foaf:name, "Natasha Noy"), (owl:sameAS, d2:nn), (foaf:name, "Natasha Noy"),(sioc:creator_of, d2:post2),(vivo:teachingOverview, "...")}*

Additionally, we make a specification of $descm(e, X)$ and define $descmp$ to project only the predicates (instead of the predicates and values as in $descm(e, X)$).

$$descmp(e, X) = \pi_{(p)}(descm(e, X)) \tag{4.4}$$

To identify the vocabularies used in the description of an entity we define:

$$vocabd(e, X) = \{home(p) | (p, o) \in desc(e, X)\} \tag{4.5}$$

4.4.2.3 Connectivity view

This view focuses on the quadruples that state the connections between the entity e and other entities. Note that this view is a subview of the description view. Here, we ignore the quadruples about e, with literal values and quadruples describing identical entities to e.

We define the entity connectivity of an entity e in $X \subseteq D_c$ as the set containing the entities targeted from e:

$$econn(e, X) = \pi_o(\sigma_{s=e}(X)) \tag{4.6}$$

Analogously, we define the dataset connectivity of an entity e in $X \subseteq D_c$ as the set containing the datasets targeted from e:

$$dconn(e, X) = \{home(o) |\ o \in econn(e, X)\} \tag{4.7}$$

Example 5 *In Table 4.1(V), econn(d1:nn,D_1)={dbr:Google,d2:p1,d3:p5,d4:p1}, whereas dconn(d1:nn,D_1)={dbr,d2,d3,d4}*

4.4.2.4 Measuring the principles at an entity and dataset level

Now that we have defined the sets for the description and the connectivity views (Section 4.4.2.1, let us look at the measures that are interesting to be applied on these sets, in order to state the extent to which the links going out of entity e follow principles P1, P2 and P3. We use the notation S to refer to any of the sets above.

Measure size Measuring the size of data is a standard way of characterizing data. We measure the size of each of the sets above by calculating the cardinality of the corresponding set (i. e. |S|).

Measure diversity When we observe if entities get their description (i. e. $classif(e, X)$ and $descm(e, X)$) extended when considering the links, we aim to identify redundancy. Furthermore, when we analyse the targeted entities and datasets, as well as the vocabularies used in the description and the links, we want to measure diversity both without and with links. In these two situations, we may encounter repetitions in the classification, the description, the entity connectivity, the dataset connectivity, and the vocabularies used in the description. Therefore, we extend the notion of our sets and model multisets (allowing repeated elements), counting the number of times each element appears in the multiset: (S, n) where n is $n : S \mapsto \mathbb{N}_{\geq 1}$, a function that given an $s \in S$ tells the number of times that s appears in S.

Diversity is a measure that takes into account the number of different (and non redundant) types of elements in a set, and at the same time takes into account how equally distributed the elements of each type are present in the set. For these two purposes, we use the Shannon Entropy (Shannon, 2001), a standard measure used in Information Theory to measure diversity.

$$H(ELS) = -\sum_{s \in S} prob(ELS = s) \times \log prob(ELS = s) \tag{4.8}$$

A low entropy value means that there is little diversity in the data. Note that $H(x) \geq 0$. In $classif(e, X)$, and $descm(e, X)$ repeated statements appear only when we consider the quadruples of the target datasets, because in one dataset quadruples are supposed to be unique. Still, we calculate entropy to be able to signal redundancy when we compare the description with and without links.

Compare measurements In order to accomplish our task of comparing measurements considering the links vs. not considering the links, we differentiate between the total set of quadruples in D_c, and the set of internal quadruples defined as:

$$D_c^{\text{internal}} = D_c \backslash I_c$$

We compare a measurement on D_c vs. the measurement on D_c^{internal} by subtracting the latter to the former.

Based on these three rationales, we define the following list of measures (cf. Table 4.2) to analyse the way links follow the principles. To measure the extension in classification, description, entity connectivity, dataset connectivity and the increase in the number of vocabularies employed, we use the difference in entropy. For example, to check if the classification is extended, we define two random variables CS (in D_c^{internal}) and CS' (in D_c) and calculate $H(CS') - H(CS)$. The difference is zero when there is no information gain, negative when redundant information is gained, and positive otherwise.

4.5 Empirical Analysis and Measure Validation

To demonstrate the feasibility of our approach for profiling the quality of links in the Linked Open Data cloud, we have implemented the measures in the SeaStar framework, which uses Java, the NxParser to parse N-Quads, and Jena for handling RDF data[5].

[5]Source code: `https://github.com/criscod/SeaStar`

ID	Principle/Description		Vars.	Definition
m11a	P1	#classes	-	$\|classif(e, D_c^{\text{internal}})\|, \|classif(e, D_c)\|$
m11c	P1	Classification Extension (entropy)	CS, CS'	$H(CS') - H(CS)$
m12a	P1	#predicate-objects	-	$\|descm(e, D_c^{\text{internal}})\|, \|descm(e, D_c)\|$
m12c	P1	Description Extension	DE, DE'	$H(DE') - H(DE)$
m13a	P1	#predicates	-	$\|descmp(e, D_c^{\text{internal}})\|, \|descmp(e, D_c)\|$
m13c	P1	Predicate Description Extension	DEP, DEP'	$H(DEP') - H(DEP)$
m21a	P2	#targeted entities	-	$\|econn(e, D_c^{\text{internal}})\|, \|econn(e, D_c)\|$
m21c	P2	Entity connectivity Extension	EC, EC'	$H(EC') - H(EC)$
m22a	P2	#targeted datasets	-	$\|dconn(e, D_c^{\text{internal}})\|, \|dconn(e, D_c)\|$
m22c	P2	dataset connectivity Extension	DC, DC'	$H(DC') - H(DC)$
m31a	P3	#Vocabularies in desc.	-	$\|vocabd(e, D_c^{\text{internal}})\|, \|vocabd(e, D_c)\|$
m31c	P3	Increase #Vocabularies Used (entropy)	VD, VD'	$H(VD') - H(VD)$

Table 4.2: List of measures to analyse the fulfilment of data interlinking principles. Columns show: the name of the measure, the principle the measure belongs to, the random variables defined for the measure, and the formal definition of the measure.

4.5.1 Data

We use data from the Linked Open Data Crawl[6], as it has been recognised as a sound snapshot of the LOD cloud in 2014 (Schmachtenberg et al., 2014). First we extracted the links from the crawled data, by parsing the dump line by line, and identifying each quadruple containing a subject and an object with different graph provenance, and therefore a different $home(x)$. While parsing the dump file, we excluded all syntactically invalid quadruples to work with clean data. Second, in order to analyse the links on a dataset basis, we split the data crawl into individual datasets, taking as contexts the dataset identifiers provided by Schmachtenberg et al.[7]. We selected a set of 35 datasets from the LOD2014 crawl (from different domains and containing several types of links), analysing a total of 1+ million links.

4.5.2 Methodology

We computed each of the measures listed in Table 4.2 for each of the linked entities in the datasets, for all types of links in the 35 datasets. Once we had all the results, we first empirically validated the measures (Section 4.5.3). After that, we analysed the results on a dataset basis (Section 4.5.4). We have published our experimental data and sources[8].

[6] Linked Data Crawl `http://goo.gl/lqxdgo`

[7] List of datasets`http://data.dws.informatik.uni-mannheim.de/lodcloud/2014/ISWC-RDB/tables/datasetsAndCategories.tsv`

[8] Experimental data and analysis code: `https://github.com/criscod/SeaStar`

Measures	m11	m12	m13	m21	m22	m31
m11	1.0	-0.24	-0.23	-0.25	-0.29	-0.23
m12		1.00	0.48	0.45	-0.34	0.70
m13			1.00	-0.19	-0.26	0.83
m21				1.0	0.94	0.05
m22					1.0	-0.08
m31						1.0

Table 4.3: Correlation between measures, for all datasets and all types of links.

4.5.3 Measure validation

Following standard practices in the literature of quality measures (Behkamal et al., 2014), we validate our measures by (i) checking that they do not provide the same measurement for all datasets D_i; and (ii) verifying that our measures are not all correlated with each other – otherwise having multiple measures would be of limited utility.

Discriminative Measures We computed for each dataset standard summary statistics such as the mean, standard deviation and quartiles considering all types of links simultaneously. As we see in the data files, the values for the measures vary across datasets. Hence, the measures are discriminative.

Independent Measures We computed the Spearman correlation of all the measurements within each dataset, putting all types of links together. Table 4.3 shows the correlation values. Measures m21 and m22 are highly correlated (0.94), which makes sense, since m21 looks at the number of target entities and m22 at the number of target datasets. In theory, one may link to many target entities within a few datasets and viceversa; but the empirical analysis suggests that having both might not particularly interesting. Having only m21 seems to be sufficient in the context of these datasets. The other two measures with a higher correlation (0.83) are m13 (predicate description extension) and m31 (increase in the number of vocabularies used).

4.5.4 Results

Let us first look at the types of links that exist in the datasets and second, at the adoption of the 3 core principles. We focus on identity links (e. g. `owl:sameAs`), relationship links (e. g. `wgs84:location`), classification links (e. g. `rdf:type`), similarity links (e. g. `skos:closeMatch`), and other more general links (e. g. `rdfs:seeAlso`).

4.5.4.1 Basic Descriptive Statistics

When we look at the type of links that is used the most in each of the datasets, in 17/35 datasets the type used at most is classification links (c), in 12/35 datasets it is relationship links (r), in 3/35 it is identity links (i) and in 3/35 it is other links (o). None of the datasets has similarity links (s). Table 4.4 shows the number of each type of link for each dataset.

4.5.4.2 Principle-based Measurements

Since our user is a data publisher willing to improve the interlinking, for each measure we analyse the inequalities among entities of the same dataset. To this end, we generate multiple box plots (one per entropy-based measure and type of link). If a box plot suggests that there are entities that get their description less extended than

Typelink	I	S	R	O	C	All
AEMET	0	0	96	0	57	153
BFS	1063	0	9	0	2862	3925
Bibbase	0	0	456	1401	0	1857
Bibsonomy	35646	0	2180	0	123080	160906
BNE	58	0	0	0	221	279
DNB	3577	0	8711	2278	55	14621
DWS Mannheim	71	0	296	39	926	1332
Eurostat	1182	0	2	0	1012	2196
Eye48	1	0	244	0	490	735
Fao	0	0	6	0	23	29
FigTrees	2	0	22	2	59	85
GeoVocab	11455	0	1759	113	7565	20892
GovWild	0	0	1998	0	0	1998
Harth	76	0	344	456	30	906
Icane	20	0	25	30	19	94
IMF	243	0	3	0	377	623
Korrekt	0	0	1174	0	7959	9133
L3S	1059	0	2478	1028	1089	5654
LinkedGeoData	634	0	12	0	254	900
LOD2	26	0	282	50	180	538
NDLJP	1	0	178	60	267	506
Ontologi	0	0	5686	0	736	6422
Openei	6	0	323	0	203	532
Reegle	327	0	432	0	135	894
Revyu	1402	0	2145	1806	39772	45125
RodEionet	9	0	981	0	0	990
SemanticWeb	2023	0	13473	421	43136	59053
Sheffield	161	0	783	0	576295	577239
Simia	121	0	2189	1	27064	29375
Soton	6691	0	25113	0	38069	69873
SWCompany	50	0	352	0	160	562
TomHeath	7	0	34	4	6	51
Torrez	0	0	266	0	493	759
TWRPI	2	0	12	0	65	79
UKPostCodes	1	0	7	0	1	9

Table 4.4: Different types of links in the 35 datasets that we analysed.

other entities in the dataset, the data publisher could think of generating further links from those entities to new target datasets. The important features of these plots are the medians (in green), the range and interquartile range—which can show big differences among the measurements of different entities when they are big—and the outliers, which in our case are relevant as they can be one of the weak spots to be improved. Compared to the original paper, some measurements have been updated.

Classification: the difference in classification cardinality considering all types of links (m11a) is 0 for 10 datasets, while for 25 datasets is 1. In the case of m11c measurements for links of type i, 25 datasets show a median of 0, and 10 show a median between 0 and 5. For links of type o and r, the m11c medians are 0 for all data sets—which makes sense, because of the definition of the measures. For links of type c, 30 datasets have a median of 0, while the rest have a median between 0 and 1 for m11c. Moreover, given the number of links of type c, we see that data publishers do also classify their entities with external classes.

Description: according to the m12a measurements, in all but three datasets the median of (p, o)-s gained is equal or below 2; the remaining datasets show a median of 4, 5 and 7. The median of new o-s gained instead (m13a) is 1 for 32 of the datasets (the other three have a median of 0, 2, and 3). Observing the m12c measurements, we notice that in links of type i (Figure 4.1) the medians vary among datasets between 0 and 8. For this type of links, there are more outliers than in classification links (see the case of Bibsonomy). It makes sense that

entities are not described homogeneously, and often publishers do not have the resources to review each generated identity link. Both things motivate that SeaStar shows the user source entities and other datasets as more positive references. The m12c measurements for links of type c (Figure 4.2) show medians between 0 and 1. For links of type r, medians range between 0 and 2; while for links of type o it ranges between 0 and 3. In the case of m13c measurements, and for all types of links, we find datasets that have negative values for the difference in entropy. That means that the links add some redundancy by adding statements with predicates that were already in the source entity. However, only a few datasets have the box in the negative area, and that happens for links of type relationship (r) (Figure 4.4) and others (o) (Figure 4.5). For example, that occurs when the data publisher adds multiple `rdfs:seeAlso` internal and external links. Comparing the box plots for identity links (type i) of the m12 and m13 measurements, we notice that in the former the range of the boxes is larger than the boxes in m13 measurements.

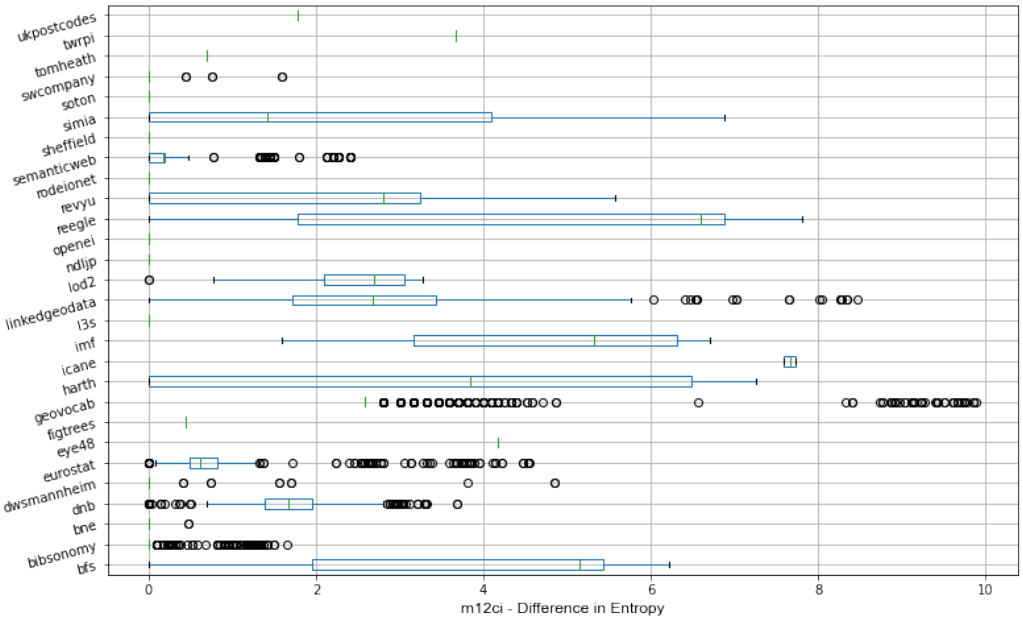

Figure 4.1: Box plot showing m12c measurements for all datasets for links of type i.

Connectivity: the medians for the number of new entities targeted (m21a) for two datasets are equal to 4, for one dataset is equal to 3, and for the rest these are all equal or below 2 new entities targeted. In the difference of entropy (m21c), the box plots do not show redundancy, which would only be possible if we compared D_c with a basis of previously generated links and new links were added over the same target entity. This would be a positive thing, if those links managed to extend the description (P1). The box plot with links of type i, shows a more skewed box (either to the left or to the right) than m12c measurements of the same type of links. M21c measurements show medians between 0 and 8 as for links of type i, between 0.0 and 3.0 for links of type r, between 0 and 2 for links of type o, and between 0 and 1 for links of type c. Regarding m22a measurements, 8 datasets show a median of 2, while the rest show a median of 1. The medians of m22c measurements are between 0 and 3 for links of types i and r, between 0 and 2.5 for links of type o, and between 0 and 1.5 for links of type c.

Vocabulary Heterogeneity: measurements m31a show that 29 datasets gain 1 vocabulary in their description, while the rest do not gain any new vocabulary. The difference in entropy (m31c) is in several datasets negative (in outliers and in the interquartile range). For links of type i the medians in m31c measurements are between -1.0 and 2.5, while for links of type c the medians are between 0.0 and 1.0. For links of type r medians are between -0.4 and 1.0, and for links of type o the medians are between -0.2 and 0.6.

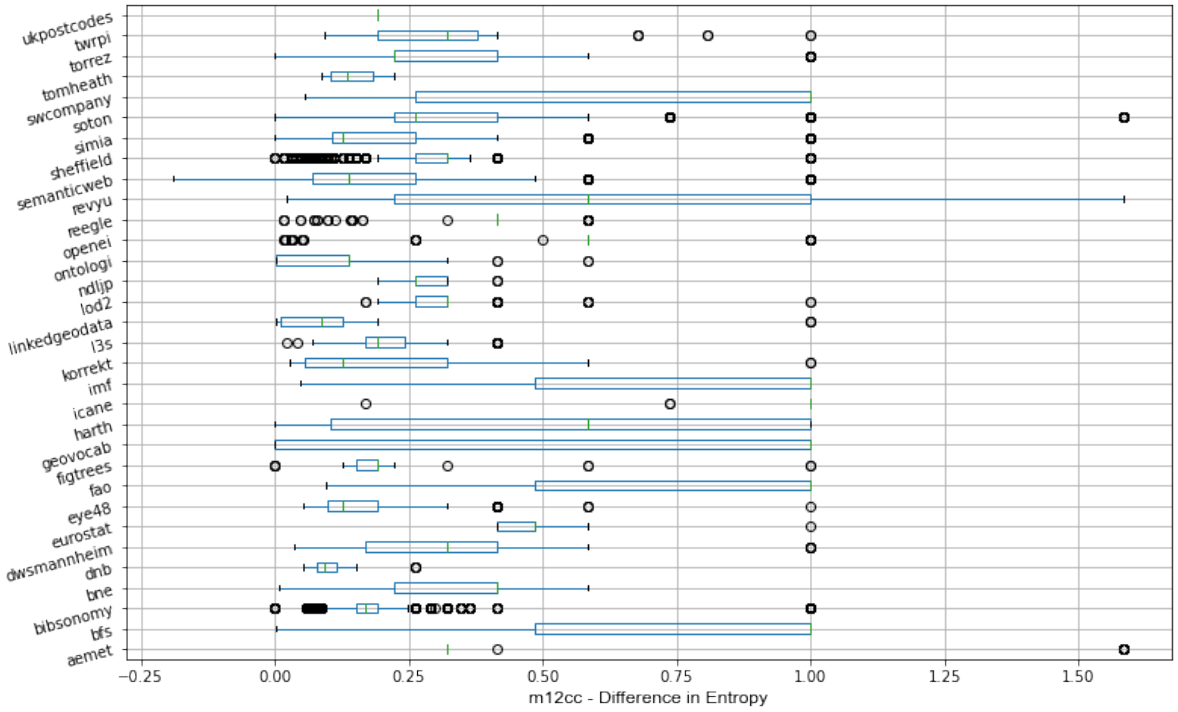

Figure 4.2: Box plot showing m12c measurements for all datasets for links of type c.

4.6 Related Work

With the growth of Linked Data, there has been an increasing interest in assessing and monitoring the quality of available data (Zaveri et al., 2016).

Status of the Linked Data Web: while there were previous studies about the conformance of the Linked Data principles (Hogan et al., 2012), the work by Schmachtenberg et al. (2014) is the most recent study on the current adoption of Linked Data best practices. With regard to the linking principle, their analysis on data crawled from 1041 distinct datasets) showed descriptive statistics about the in- and out-degree of datasets (defined by the number of datasets pointing to / targeted by the datasets), and the most frequently used predicates.

Link Analysis: there are studies focusing exclusively on links. Halpin et al. (2010) analyzed the usage of the `owl:sameAs` predicate in the links of the Linked Data space. They observed that sometimes the predicate was used with a meaning different from its original definition, and suggested to improve the quality of such links by using alternative and more suitable predicates (e. g. `skos:closeMatch` when not all properties of the target entity apply to the source entity; `foaf:primaryTopicOf` when the target entity represents but is not the same as the source entity). Hu et al. (2015) empirically studied term and entity links in Biomedical Linked Data. Their findings include link and degree distributions, the analysis of symmetry and transitivity, and the evaluation of entity matching approaches over the links. Neto et al. (2016) analysed the Linked Data crawl by Schmachtenberg et al., together with the set of Linked Open Vocabularies[9]. They examined the number of valid and dead links (i. e. in their work, links with an *o* that cannot be described in the target distribution), as well as the number of namespaces in link distributions and datasets. Albertoni and Pérez (2013); Albertoni et al. (2015) analysed the completeness of the interlinking of pairs of datasets and the extent to which datasets become more multilingual thanks to the links. These methods fail in stating the extent to which links add value to the source dataset in terms of the principles that we mention in this paper.

[9]LOV `http://lov.okfn.org/dataset/lov/`

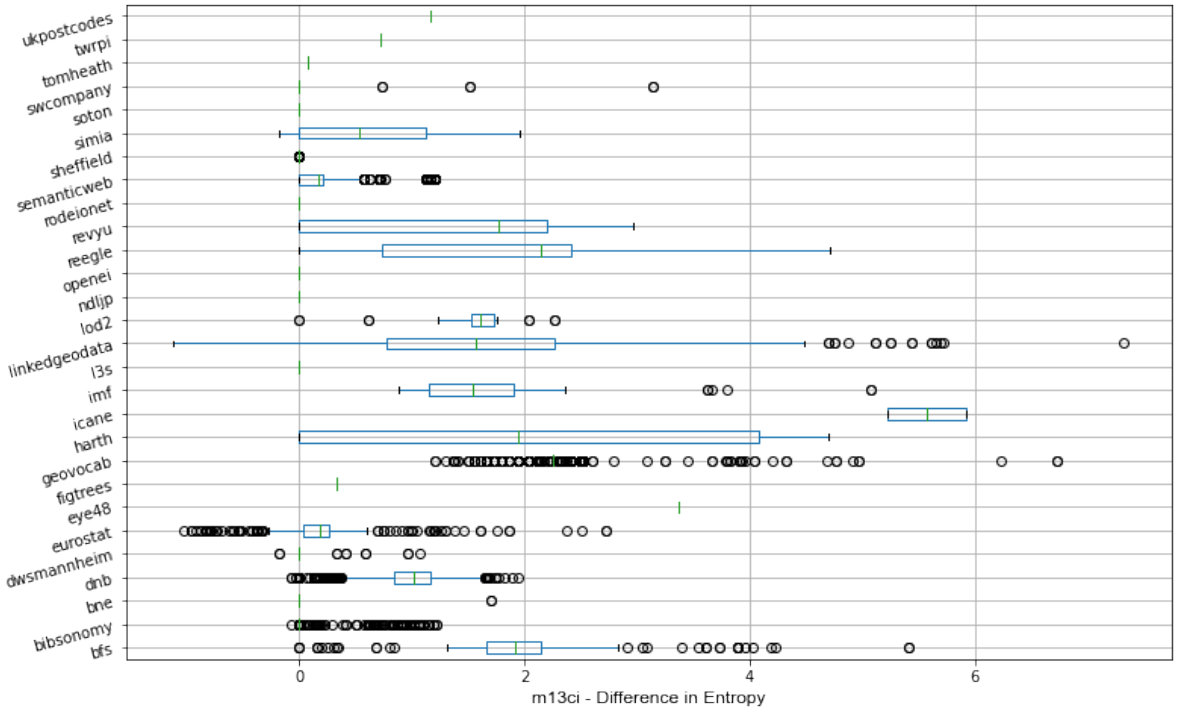

Figure 4.3: Box plot showing m13c measurements for all datasets for links of type i.

Methods for Assessing Link Quality: several methods have been developed to assess the semantic accuracy of links (e. g. to decide whether `ch:koblenz owl:sameAs de:koblenz` holds or not). Guéret et al. (2012) defined a framework including three measures from the area of network theory: degree, clustering coefficient and betweeness centrality of the entities in links; as well as two measures that the authors define: number of unclosed same as chains and description enrichment defined as the raw number of new statements gained by the source entity. While Guéret's et al. notion of description enrichment is related to ours, the main differences are that we are able to observe further dimensions (e. g. how the classification of entities and the connectivity is extended by the links), our approach is not only restricted to `owl:sameAs` links (as it applies to any link) and we are able to signal redundancy.

4.7 Conclusions and Future Work

We have presented a collection of measures whose goal is to help in gaining insights into the quality of existing links, and understanding the effect that links produce in the source dataset. After analyzing 35 datasets of the LOD cloud with these measures our findings show that source entities are classified with external classes via links, and identity links contribute to inheriting new classes. We also observed that there is certain redundancy in the properties and vocabularies used as for extending the description. The differences between entities and datasets shown in the boxplots justify the need for our framework, which is able to pinpoint inequalities in link-based information gain, as well as reference interlinked entities and datasets to data publishers.

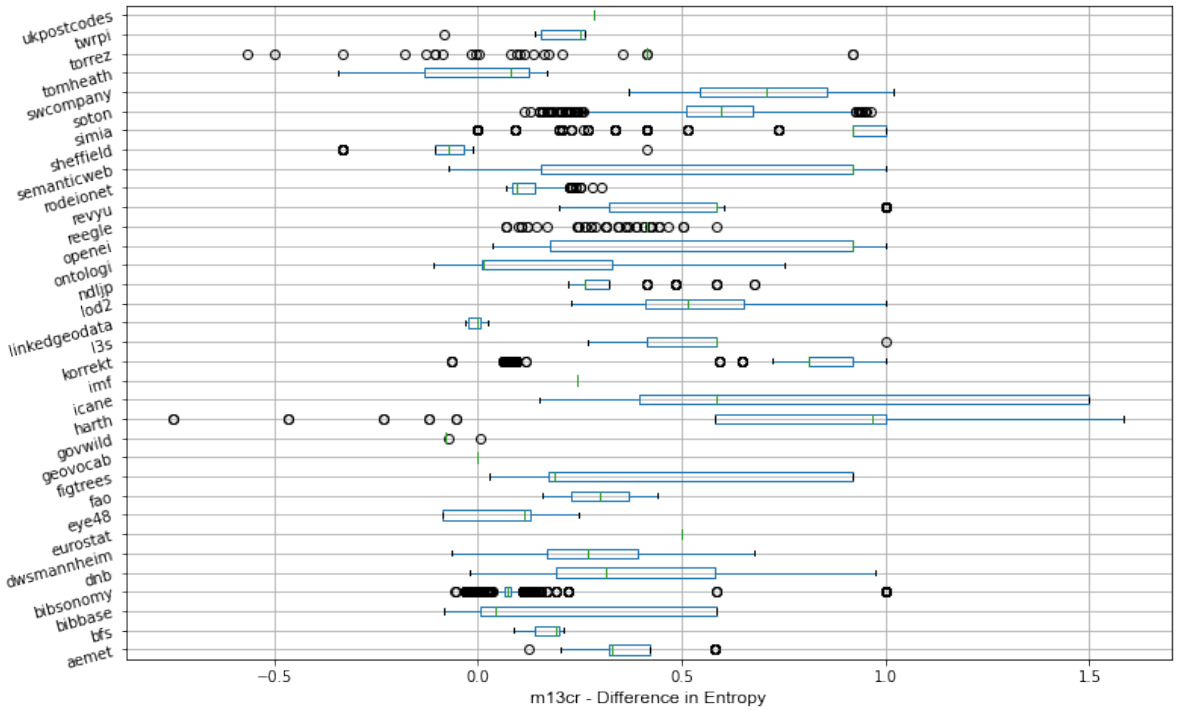

Figure 4.4: Box plot showing m13c measurements for all datasets for links of type r.

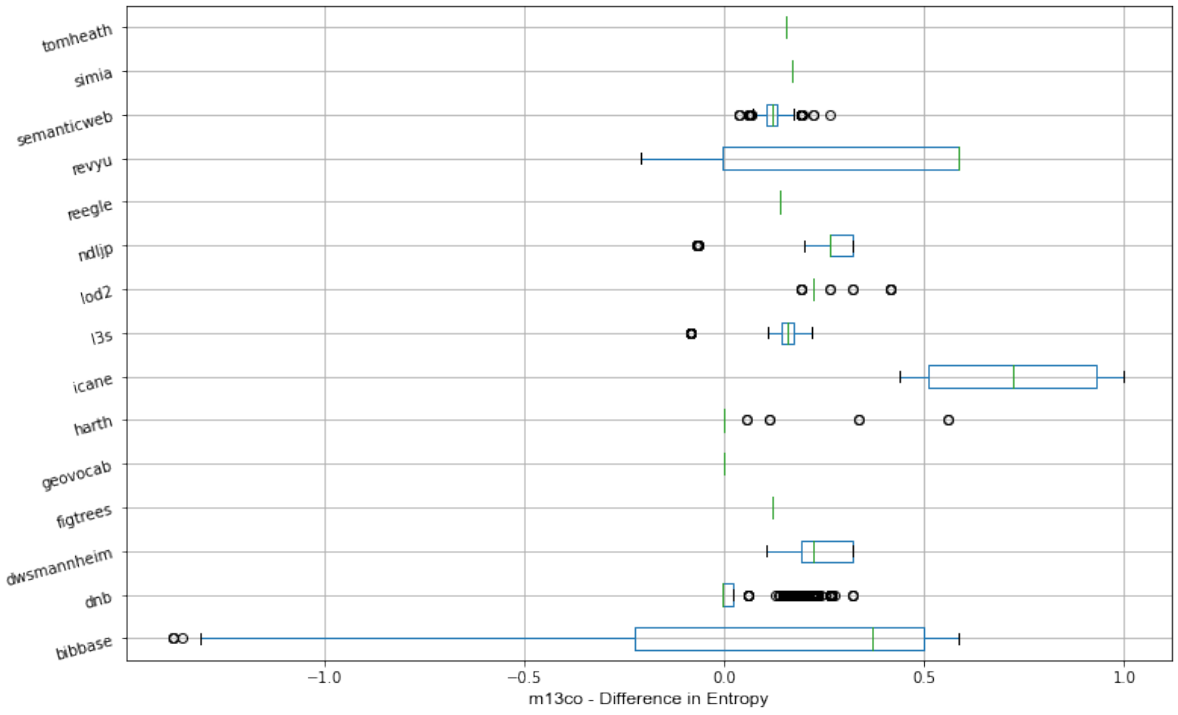

Figure 4.5: Box plot showing m13c measurements for all datasets for links of type o.

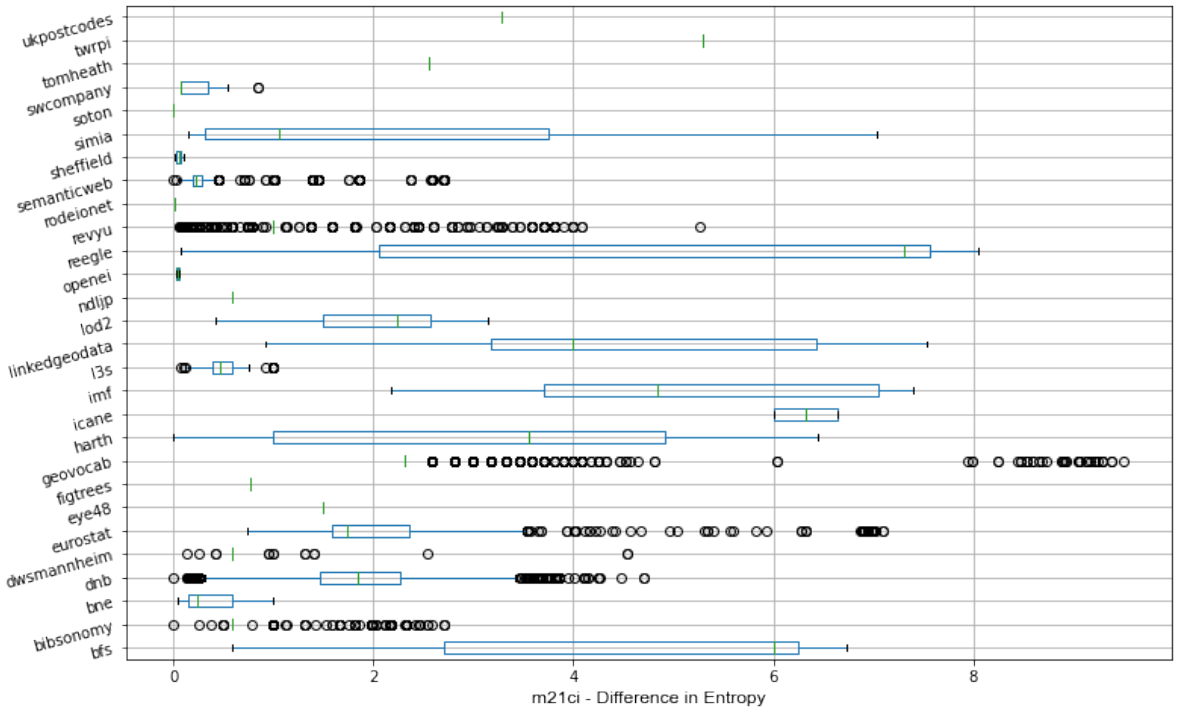

Figure 4.6: Boxplot showing m21c measurements for all datasets for links of type i.

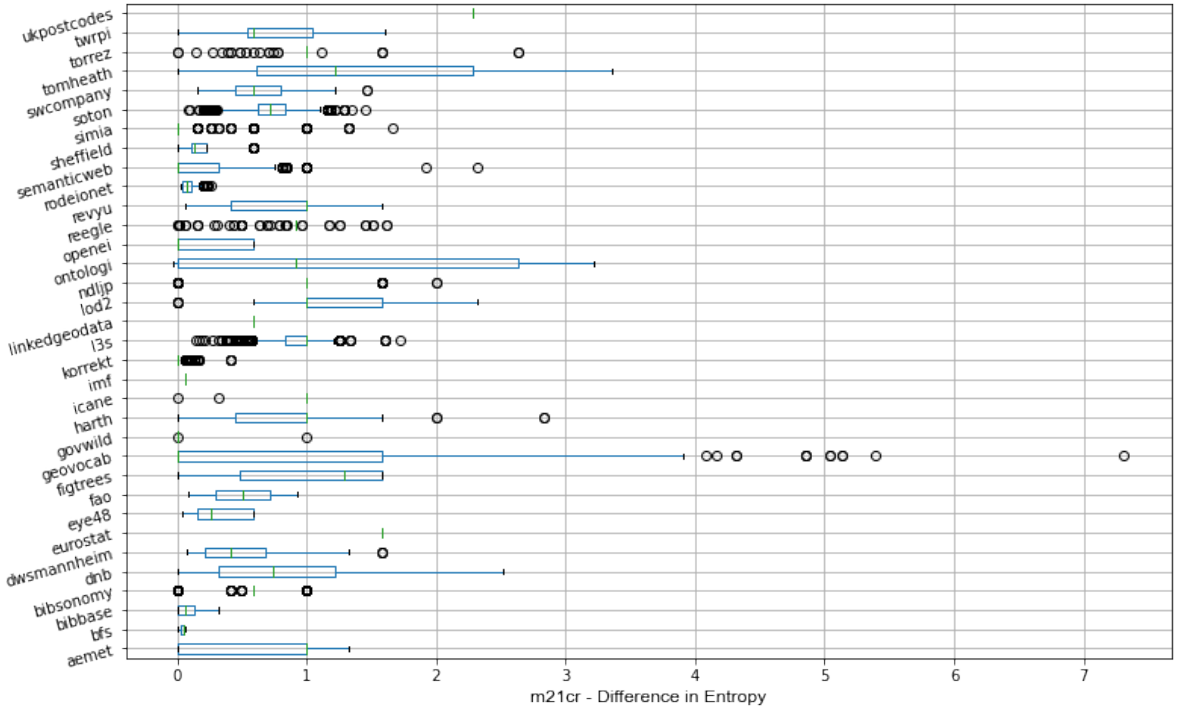

Figure 4.7: Boxplot showing m21c measurements for all datasets for links of type r.

Chapter 5

The Evolution of Power and Standard Wikidata Editors: Comparing Editing Behaviour over Time to Predict Lifespan and Volume of Edits

Abstract: Knowledge bases are becoming a key asset leveraged for various types of applications on the Web, from search engines presenting 'entity cards' as the result of a query, to the use of structured data of knowledge bases to empower virtual personal assistants. Wikidata is an open general-interest knowledge base that is collaboratively developed and maintained by a community of thousands of volunteers. One of the major challenges faced in such a crowdsourcing project is to attain a high level of editor engagement. In order to intervene and encourage editors to be more committed to editing Wikidata, it is important to be able to predict at an early stage, whether an editor will or not become an engaged editor. In this paper, we investigate this problem and study the evolution that editors with different levels of engagement exhibit in their editing behaviour over time. We measure an editor's engagement in terms of (i) the volume of edits provided by the editor and (ii) their lifespan (i. e. the length of time for which an editor is present at Wikidata). The large-scale longitudinal data analysis that we perform covers Wikidata edits over almost 4 years. We monitor evolution in a session-by-session- and monthly-basis, observing the way the participation, the volume and the diversity of edits done by Wikidata editors change. Using the findings in our exploratory analysis, we define and implement prediction models that use the multiple evolution indicators.

5.1 Introduction

Knowledge Bases have become key to enabling semantic search and exploration functionalities in a wide range of applications on the Web. Besides the general interest knowledge bases owned and managed by companies (e. g. Google's Knowledge Graph), there exist openly available knowledge bases as a result of open data initiatives. For example, the Linked Open Data (Schmachtenberg et al., 2014) consists of over one thousand structured datasets describing various topical domains, and published by different agents including universities, private companies and other organisations. Wikidata is an open, free, multilingual knowledge base (Vrandecic and Krötzsch, 2014) started by Wikimedia Deutschland, containing as of October 2017 more than 37 million data items. While the creation of Wikidata was motivated by the need of more efficient data management methods within Wikipedia, Wikidata has become an important knowledge base for many other systems and applications that reuse Wikidata's item descriptions. Moreover, hundreds of other external datasets such as VIAF, the Library of Congress, Europeana or Facebook Places created by libraries, governmental and private organisations have been integrated with Wikidata, making Wikidata the hub to explore a network of open knowledge spread over the Web.

Wikidata, in contrast to the vast majority of knowledge bases on the Web of Data, has a strong focus on human intervention across its complete data management process. Tools and bots operate on the data to some extent, but the data is primarily curated and maintained by a community of volunteers. The community – editors, developers, data providers and researchers – discusses and collaborates to decide how to model, ingest and patrol information. This human-driven process enables Wikidata to diminish the data quality problems that appear in (semi-)automatically generated knowledge bases, including entity misclassification, inconsistencies, semantically inaccurate links, and outdated data (Zaveri et al., 2016).

Weaving a community of devoted people who are able to contribute duly of their own free will is one of the main challenges in Wikidata. Since Wikidata's mission is to represent human knowledge in a structured way, the project needs the help of a large number of people. Despite the positive response from thousands of volunteers, there is a clear need for attracting new contributors and growing the community, because there is still much data and many tasks to work on [1]. At the same time, it is important to retain contributors who already approached Wikidata by encouraging them to contribute. To address this challenge of acquiring new people and maintaining the activity of existing community members, Wikimedia organises events to advocate for free knowledge, disseminate the project, offer technical training for newcomers, establish social tights between members, as well as edit data and develop software together. There are also resources that aim to facilitate the editors' contribution. For example, the Wikidata Games [2] present editors with very simple edit requests (e. g. add the occupation or the gender of a person) that help to improve the completeness of the knowledge base. Still, Wikimedia's official reports indicate that there is a big number of the people who one day edited Wikidata but are currently inactive. The latest statistics [3] show that from August 2016 until August 2017 the number of active editors [4] was between 7.8K and 8.7K. Taking into account that Wikidata registered more than three hundred thousand contributors during its history, the inactivity ratio is large. This fact could be detrimental for Wikidata, especially when we consider that many of these inactive contributors did not edit for a very long time and might have, presumably, abandoned the project. This level of abandonment rate suggests that Wikidata may not be maintained nor extended to its full potential. One may think that increasing editors retention in Wikidata could lead to more item descriptions and more complete and up-to-date data.

In order to improve the situation and reduce editor attrition, the community (and especially Wikimedia as the main manager) needs to design and implement methods that stimulate a change in the behavior of these contributors who become non-active. The literature in marketing research has extensively studied the problem of customer churn (or customer turnover) and designed churn management strategies that aim at increasing customer retention (Rosenberg and Czepiel, 1984; Ang and Buttle, 2006), because losing customers can endanger a business – having fewer customers often means obtaining a lower revenue. While in Wikidata (and in general in any Wikimedia project) the motivation to keep editors contributing to the knowledge base is not economic, there is attrition and participation inequality. Therefore, even though the concrete actions to engage further contributors are different from scenarios with economic transactions and customers, it makes sense to use assumptions and techniques from this field of research.

Marketing analysts and researchers highlight the importance of targeting customers to improve retention (Verhoef, 2003). However, it is important to provide tailored solutions based on the customers' behavior, especially since loyal customers need a different attention than likely-to-drop-out customers. That is why traditional retention strategies first tend to predict whether a customer will be a *churner* or not [5], and then implement actions to convert likely-to-be churners into non-churners (Gordini and Veglio, 2017; Difallah et al., 2014). For these actions to work effectively, the prediction needs to be done as early as possible. The subfield of survival analysis (Cox, 1992), often used in the context of retention management, estimates the "time left" until an event (e. g. time until death in the medical domain, time until drop out in the context of online shopping). This is useful

[1] Wikidata's Phabricator Ticketing System `https://phabricator.wikimedia.org/tag/wikidata/`

[2] Wikidata Game `https://tools.wmflabs.org/wikidata-game/distributed/#`

[3] Wikidata Revolution presentation at Wikimania 2017, in August 2017 `https://wikimania2017.wikimedia.org/wiki/Submissions/The_(Wiki)Data_(R)Evolution`

[4] According to Wikimedia Foundation, an active user is "A user with 5+ edits in the main namespace of a given project over the last 30 days" `https://meta.wikimedia.org/wiki/Research:Metrics`

[5] Retention Science `https://go.retentionscience.com/hubfs/Documents/Retention_Science_Predicting_Customer_Churn_Guide.pdf?t=1507690636756`

to act against retention and intervene before customers leave. Hence, having a method to estimate the time that customers will spend in the system is a prerequisite to design a retention management solution.

In this article, we work towards developing such a method in the context of Wikidata. We aim at predicting the time that contributors will be in Wikidata – lifespan, as well as the volume of edits they will contribute with, to understand the magnitude of their action in the knowledge base. Therefore, our goal in this work is not to provide the solution to the retention problem designing a retention campaign but to build the prerequisites, providing a data-driven approach to the prediction task. In order to do so, first we carry out an exploratory analysis to understand differences between power and standard editors, and second, we provide and evaluate a predictive model that uses the insights of the exploratory data analysis.

Given that we look at lifespan and volume of edits, we consider 4 types of contributors: (a) contributors with *long lifespan* and *high volume* of edits, (b) contributors with *short lifespan* and *high volume* of edits, (c) contributors with *long lifespan* and *low volume* of edits and (d) contributors with *short lifespan* and *low volume* of edits. The contributors with the highest impact on the system are contributors who contribute extensively, and for a long time (group a). Groups (b) and (c) are also valuable contributors. For example, a contributor supervising recent changes to revert and correct malicious edits, a few times a month; she might do only a couple of edits a month, but if she does it for a long period of time, her contribution can help Wikidata in terms of data quality. Yet, ideally one would like to have as many contributors as possible in group (a). Predicting the lifespan and volume of the contribution of editors, we are able to classify existing contributors into one of these groups, and we can consequently, decide whom to address and how to do it.

In order to make such a prediction, we analyze the evolution of editing behaviour. We analyze this information from two different perspectives: we run a **(i) session-based analysis**, and we also study the editing progress **(ii) on a monthly basis**. The two perspectives are complementary: with the former, we aim at understanding the extent to which editors change their behaviour as they gain more experience and do more edits in each session they spend editing; while with the latter, we perform a time-sensitive analysis. When we analyze the behavioural evolution throughout the editors' lifetime (in sessions and in months) we measure indicators related to the editors' productivity, editors' participation and the diversity of the types of edits (cf Section 5.7).

To the best of our knowledge, there is no previous work analyzing the evolution of editing behaviour in these terms as a predictor of the volume of edits and lifespan in Wikidata. The research around Wikipedia, older than Wikidata, has examined the edit history in terms of edit quality, editor interaction, editor participation, as well as emerging information cascades (see Section 5.3 for a detailed description of the Related Work). Even if both systems share commonalities, Wikidata has features that could, in theory, encourage people to work with different patterns than in Wikipedia. Besides that, there is no published work about the intersection of both communities; so, it should not be assumed that contributors in Wikipedia and Wikidata work exactly in the same way. It is, thus, important to collect empirical evidence and study the behaviour of contributors in the Wikidata environment, too. Anyhow, other works have not used the trend in the evolution of contribution, participation and diversity as predictive factors of lifespan and volume of edits. Therefore, not only do we contribute to the state of the art by studying contributors in Wikidata, but we also contribute with a method. In the context of Wikidata, Müller-Birn et al. (2015) grouped editors based on the types of tasks they focus on, and as an extension (Cuong and Müller-Birn, 2016) observed the extent to which Wikidata editors change between roles. Piscopo et al. (2016) surveyed Wikidata editors to understand differences between novice and expert users, examining how motivations, goals, usage of interfaces and type of actions differ. While these works help to understand the fundamental differences between some groups of editors, and the way editors change the type of actions they work on as they become more experienced, these works fail to (i) provide a method that helps to predict the extent to which editors will be engaged in terms of time and contribution, and they (ii) do not explain how power and standard editors change or maintain their contribution, participation, and diversity of type of edits.

The main contributions of this paper are:

- We run a quantitative analysis about the volume of edits and the lifespan of editors (Section 5.6).
- We perform a longitudinal study over the Wikidata history and identify the trends in the evolution of editing behavior of different groups of editors, mainly standard and power editors (Section 5.7).
- We define supervised classification methods to predict the range of months that an editor will be contributing to Wikidata, and the range of edits that an editor will do in Wikidata (Section 5.8).

- We highlight a set of implications that our findings may have in the Wikidata community Section 5.9.

5.2 Wikidata: A Crowdsourced Knowledge Base

Wikidata is a freely-available openly-editable knowledge base. It is the result of a continuous community effort started by Wikimedia Germany in 2012. More than just human contributions, Wikidata serves as a data integration hub where other knowledge bases (e. g. VIAF, Europeana, DBpedia) link to or are imported in[6].

Wikidata is set apart from many of the other available knowledge bases in several ways:

- *Community curated*: Wikidata's initial primary goal was to support Wikipedia editors by providing them with a central knowledge base that holds data to be shared between all Wikimedia projects. In order for this to happen and in the spirit of its sister projects, Wikidata is open for editing by anyone and maintained by an open community of editors.

- *Multilingual*: In order to fulfil its initial primary goal of supporting Wikipedia editors, Wikidata needs to provide one central place to collect and maintain the same data that is then shared between all Wikimedia projects. All editors must work on the same data independent of their language. Wikidata achieves this goal by identifying its items and properties (the basic building blocks of Wikidata used to describe concepts in the real world) with language-neutral identifiers. For example, the property "instance of" is identified by "P21" and the concept "Earth" is identified by "Q2".

- *Knowledge diversity*: Wikidata is built as a secondary database. This means it is not meant to record raw facts. Instead, it collects statements from other sources and references them. With this model, different views can be recorded on controversial topics and the consumer of the data can investigate further and judge which of the sources they accept. This is crucial for Wikidata in order to cater to the many different cultures in Wikimedia projects. It also aligns with Wikipedia's ethos of referencing information and making it possible for the reader to dig deeper into a topic.

Wikidata is set apart from its sister project Wikipedia in several ways as well:

- *Language and culture*: Wikipedia is divided by language and by extension culture. In Wikidata editors from all these languages and cultures come together to work together on the same data.

- *Notability*: Wikidata serves all Wikimedia projects and therefore needs to cover more concepts than any of the individual Wikipedias.

- *Large scale editing*: Wikidata, as a result of its virtue of being machine-readable and editable, is seeing considerably more edits done with the help of tools and bots than the Wikipedias. Indeed, it currently accumulates one third of all edits across Wikimedia projects.

- *Editing interface*: Wikipedia offers a text-editor like interface as well as a WYSIWYG editor. Wikidata offers a form-based interface as well as a large amount of special-purpose tools (e g. WikiShootMe, Wikidata Game).

5.3 Related Work

The consolidation of social computing has led to multiple studies that examine human behaviour in various systems, including major volunteer crowdsourcing projects like Wikipedia and Open Street Maps. We review related work focusing on methods and findings about the contribution of Wikidata's and Wikipedia's volunteers, user attrition in Web systems and the evolution of user behavior in volunteer systems.

[6] In 2015 Google announced the port of the Freebase knowledge base to Wikidata.

5.3.1 General Knowledge about Volunteers' Contribution in Wikidata

One of the first works looking at Wikidata, by Müller-Birn et al. (2015), analyzed the emerging roles of editors, in terms of the possible operations in Wikidata, including the creation of items and ontology elements, as well as the addition of references. The results showed that a majority of editors have specialised contributions, and only a small active group contribute across many areas of the project. As a continuation of that work, Cuong and Müller-Birn (2016) looked at 2 years of edit history, to analyse the transitions between editorial roles and found that "users who joined earlier are persistent contributors even though they take part in different roles, whereas users who join late are quite stable in their behavior". We divide contributors in a different way, based on lifespan and volume of edits instead, and we run a temporal-based analysis.

A more recent work by Piscopo et al. (2016) looked at the way editors grow from novice to proficient contributors. The authors performed a qualitative analysis surveying Wikidata editors, and found that editors become more responsible for their work over time, and as time goes by they participate more with the community, carry out more advanced tasks and use more different tools. Compared to this work, our differences are that we have a different focus (prediction using the evolution), we look at the progress of different dimensions (contribution, participation and diversity of the type of edits), and we run a data-driven quantitative analysis considering edit sessions – while Piscopo et al. (2016) surveyed editors.

5.3.2 General Knowledge about Contributions in Wikipedia and Other Knowledge Bases

West et al. (2012) look at who the Wikipedia editors are and how their Web usage patterns differ from non-editors concluding that editors are typically more expert on certain topics and get informed on the Web before starting to edit Wikipedia. In another work looking at Wikipedia edit patterns, Yasseri et al. (2012) performed a large scale analysis of Wikipedia edit data across different languages by geo-locating editors and looking at temporal edit patterns. The authors identified two main categories of editors: those more active during week days and those more active during weekends. Iba et al. (2010) leveraged Wikipedia edit data to build a social network across editors. They observed a differentiation between two main types of articles: those topic-focused having few expert editors and those of general interest involving many casual editors.

Previous work on Wikipedia has also focused on measuring and predicting the quality of contributions. For example, Druck et al. (2008) looked at contribution quality prediction using the revert time of an edit (*expected longevity*) as quality measure. The prediction methods used are based on features such as change type, words used, article, and user. The authors concluded that the prediction of the edit quality depends very much content-dependent, and not only user-dependent. Another work looking at quality, by Halfaker et al. focused on how editors perceive revert actions on their contributions (Halfaker et al., 2011). The authors showed that reverts are good to improve the overall quality of Wikipedia but also affect user motivation and future engagement. They also highlighted the 'newcomer retention problem' which we look at in our paper. Finally, there are works like (Walk et al., 2015a) by Walk et al. that studied the way the content of the knowledge base impacts the editing behaviour. The authors studied the sequences of edits that users in an ontology editing environment did over several ontologies, and compared different hypotheses using the HypTrails (Singer et al., 2015) method. They observed that the hierarchical structure of the ontology and the entity similarity are the dimensions that have the strongest influence on the behaviour of editors.

The focus of our work is primarily on the editors' engagement patterns, independently from Wikidata's content because we would like to first understand if there are editors with different levels of commitment and different habits. This is high importance to the community, because a project like Wikidata greatly benefits from unconditional contributors who provide knowledge, no matter the status of the knowledge base.

5.3.3 User Engagement and Attrition in Volunteer Communities

Measuring user engagement in volunteer projects is key to understand who the most valuable users are and to design mechanisms that decrease attrition. A work related to the effect of Wikipedia edit activity on engagement is (Gandica et al., 2015), which defines a function of edit probability where the more a user has edited the more

likely she is to edit in the future. Danescu-Niculescu-Mizil et al. (2013) study how users join and leave an online community focusing on the linguistic aspects of their contributions. They provided a method for predicting the range of time when users would drop out of the community based on their use of words over time when writing posts in a beer-related community. Their empirical research showed that the linguistic evolution stabilises after a while staying stable until drop out. As the authors showed, in the dataset they analysed, this phenomenon is relative to the lifetime of the user, and not to absolute to a biological frame. Ponciano and Brasileiro (2014) measured the activity ratio and the daily devoted time by users in the citizen projects of Galaxy Zoo and The Milky Way Project, grouping people into profiles like the hardworking, spasmodic, persistent, lasting and moderate users. In both datasets, they found that the majority of users are classified as moderate, and only a few are persistent users.

In our work we also look at contribution and participation, but using different measures (i. e. number of edits per month / session, number of edits per item, number of items and seconds invested in the session) and observing the trend –increasing, decreasing or constant– of the measurement over time. Furthermore, we compare power and standard users, without clustering them.

5.3.4 Evolution of User Behavior in Volunteer Communities

The work of Geiger and Halfaker (2013) was one of the first works applying the common technique in Information Retrieval of grouping edit activities into edit sessions in which an editor performs a number of related edits and then stops for a certain period of time (e.g., few days or weeks). Geiger et al. empirically defined Wikipedia sessions as of a one hour inter-edit time and then looked at time elapsed between sessions and at the evolution of sessions over years. In our work we follow the methodology they used to define edit sessions in Wikidata, and use the inter-session time to observe participation. As a difference, we look at the progress over time to use it in the prediction.

The study done by Panciera et al. on Wikipedia editors (Panciera et al., 2009) is extremely relevant for our work. The authors discovered that "Wikipedians are born, not made", which means that editors do not contribute more, more frequently or with higher quality over time, and rather maintain high levels of contribution. As the authors explain, these findings suggest that the system does not encourage further engagement and people who are truly committed are devoted to the cause of free knowledge from the beginning. In our work we test this hypothesis in the context of Wikidata, adding hypotheses and observations about participation and task diversity over time, as well as defining evolution in two different ways (i. e based on months and sessions). Moreover, we complement the exploratory research about the evolution of editing behavior with a prediction problem in terms of lifespan and volume of edits, as prediction is a central component in churn management and in computational social science in general (Alvarez, 2016; Strohmaier and Wagner, 2014).

Walk et al. (2016) provided a model for activity decay in collaboration networks that captures activity decay rate and peer influence growth rate. They evaluated their approach with Semantic MediaWiki datasets, to prove that the activity dynamics they simulated is close to reality. While the goal of the paper is not to describe the datasets, the empirical evaluation showcased the activity decay present in several communities. In our works, we do not observe the interaction between editors.

In the context of OpenStreetMap and humanitarian mapping, Dittus et al. (2016) found that different contributor cohorts (working around different initiatives) show different retention patterns. Initiatives that were designed with tight coordination practices (e. g. including mapathons) had contributors with higher retention. Interestingly, the authors also found that early abandonment was related to higher contribution, suggesting that some standard contributors had a burnout effect. In the descriptive part of our research we measure the relation between lifespan and volume of edits.

5.4 Research Hypotheses

The volunteering-based design in Wikidata, akin to any other Wikimedia project, encourages the contribution of intrinsically motivated people who believe in free knowledge and are eager to help with their expertise and cooperation. The system provides mechanisms for accountability and transparency (i. e. anyone can see who did what,

and discussions are public), and strong contributions (and contributors) are openly acknowledged and recognised. The feeling of belonging to the community also drives volunteers in Wikidata, like in Wikipedia (Nov, 2007).

Because our goal is to understand and predict who will and who will not thrive as volunteer, we support our research in past studies and related theories that highlight differences in the behavioural evolution of effective and non effective people –related to the volume of edits – and committed/persistent and uncommitted people – related to lifespan.

The work by Panciera et al. (2009) suggests that Wikipedians maintain a constant level of contribution. The fact that the Wikidata and Wikipedia communities share some commonalities, makes us hypothesize that this is a key difference between power and standard editors in Wikidata, too. Note that it is still worthwhile testing the hypothesis empirically, because Wikidata has many more ways to contribute than Wikipedia, and it has a unique feature regarding knowledge curation, as compared to Wikipedia, that lays in the intrinsic structured nature of its content (i. e. each item is formed by a collection of structured factual statements rather than encyclopaedic articles written in natural language). These two differences could potentially influence editing behaviors. If editors show a constant editing behavior, it suggests that they have habits. Habits are related to commitment and effectiveness (Duhigg, 2012), and often the users who do not develop habits are related to churn. We set up three hypotheses, focusing on contribution (e. g. number of edits per session / month), participation (e. g. time invested in a session) and diversity (e. g. type of task variability).

Hypothesis 1: *A constant contribution over time is a signal of power editors but not of standard editors.*

If editors develop their editing habits, they are likely to schedule them regularly in their agendas, and the longer a habit runs for, the more established it becomes (Duhigg, 2012). So, in terms of the time spent while contributing, we hypothesise:

Hypothesis 2: *A constant participation over time is a signal of power editors but not of standard editors*

The fact that Piscopo et al. (2016) found surveying editors, that indicates that editors take more responsibility and do different tasks over time (when they grow from novice to proficient), makes us hypothesise that an increasing trend in the diversity of tasks over time differentiates power from standard editors, assuming that a standard editor will have a lower probability of crossing the line from novice to proficient editor. Hence, we formulate the third hypothesis as:

Hypothesis 3: *An increasing diversity in the types of tasks done is a signal of power users but not of standard users*

If these hypotheses are confirmed, these dimensions will help us predict the class to which contributors will belong (i. e. power or standard contributors in terms of lifespan and volume of edits).

5.5 The Wikidata Edit History Dataset

We obtained the XML data dump (as of 01.07.2016 [7]) provided by Wikimedia containing information about each of the editing actions done by contributors. We parsed the data, transformed it into CSV data, and imported into a memory-based database (MemSQL). For each edit, we kept information about the editor who did the edit, the timestamp when the edit was completed in Wikidata's database, the item where the edit was done, and the comment that MediaWiki automatically generates to annotate the changes in the database. We then classified the edits based on the (i) type of editor, (ii) the type of thing edited, (iii) the means used to edit, and (iv) the type of edit carried out.

- *T*ype of editors: We distinguish between users who are registered users (i.e. have a username and edit Wikidata while being logged in) and users who are anonymous (and from whom we only know an IP address). Registered users can be humans or bots. We identify bots by looking up the public list of registered bots and discard the edits done by this set of users, because we are primarily interested in understanding human behaviors of Wikidata editors. It is important to distinguish between registered and anonymous users, not only because people might behave differently when they reveal their identity, as Huang and Fu (2013) showed in the context of microtask crowdsourcing, but also because non-registered edits might also

[7] Wikidata Wiki dump `https://dumps.wikimedia.org/other/incr/wikidatawiki/`

come from applications implementing automatic edits via the Wikidata API, and hence show a different behaviour.

- *T*ype of things edited: We distinguish between item edits and non-item edits. Item edits are edits done to create, update or delete an item in the knowledge base (e.g. an entity, a class). Non-item edits are edits done in other kinds of pages such as project and user pages. We only focus on edits done on items which are part of the knowledge graph.
- *M*eans to edit: There are various interfaces to edit Wikidata (e.g. the wiki, Wikidata games, etc.). We differentiate between edits done using tools and edits done without tools, because in the former case users do not decide what item to work on next, nor the type of edit to do. To classify edits into these two groups we use the *O*Auth tags database provided by Wikimedia and scan the edit comments for any other trace left by tools listed in Wikimedia directories (including Gadgets. User scripts and external tools).
- *T*ype of edit: we use the list of actions registered in Wikidata's backend [8] to distinguish between the major actions (e. g. set a label, update a claim, delete a claim or add a reference).

We published our research data online [9].

5.6 Quantitative Analysis of the Wikidata Edits

Before addressing the topics raised in the definition of hypotheses, we obtain descriptive statistics that allow us gain a better understanding of the dataset. Out of the complete set of 350+ million edits, 1.5+ million edits are done by anonymous users and 261+ million edits are done by bots. We exclude both sets of edits from our analysis, as we are interested in studying human editing behaviour. We limit our analysis therefore to a raw dataset of 87+ million edits.

As expected, the number of edits done with tools exceeds the number of edits done manually (see Table 5.1). The number of distinct editors editing manually (without tools) is higher than the ones using tools. An explanation to this fact might be that there are sporadic editors who make a few edits in the wiki to try out the system or to maintain specific targets, but do not get involved with tools like the Wikidata Game or QuickStatements.

	Registered Without tools	**Registered With Tools**
Number of edits	35,069,629	52,345,356
Items edited	7,633,131	13,065,045
Non-Items edited	176,892	1,392
Number of distinct editors	142,643	6,060

Table 5.1: Different types of edits in Wikidata's history (from October 2012 until July 2016) that we consider in our analysis.

Given that editing behaviours may show different type of patterns when using tools, we decided to focus our analysis on the set of edits done by registered users without tools. The data used by default for all the results presented in all the following sections is therefore referring to this set of 35+ million edits done over 7+ million items of the knowledge base.

[8] Wikibase actions `https://www.mediawiki.org/wiki/Wikibase/API/de`
[9] Research Data `https://github.com/criscod/wikidata_editors_evolution_jcscw`

5.6.1 Volume of Edits

To understand the influence that different users have in the system, we computed the total number of edits made by each user and plotted the histogram. As it can be seen in Figure 5.1, there are many users who make a low number of edits and few users who make a high number of edits, as expected. The median of the edit counts is 2 edits, while the standard deviation is 7223 edits. This kind of behavior is actually similar to what we observe in other crowd-powered systems. In the context of paid microtask crowdsourcing, participation is typically dominated by few workers who complete most of the workload (Franklin et al., 2011). Similarly, in citizen science projects (e. g. GalaxyZoo) very few users perform many tasks, while the vast majority tags less than 30 images each (Lintott et al., 2008).

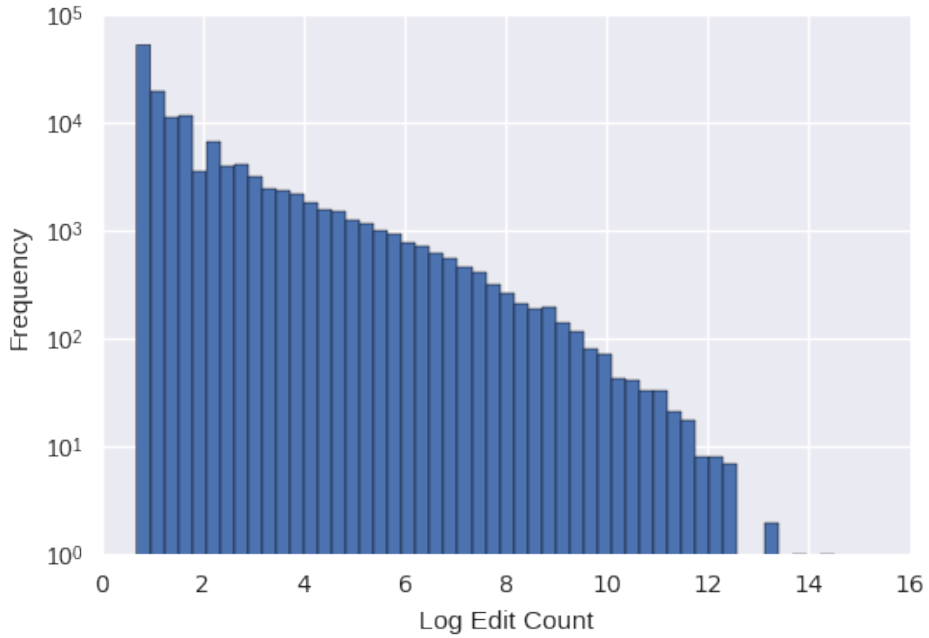

Figure 5.1: Total number of edits done by each Wikidata user.

We also looked at the extent to which items are crowdsourced, by computing the number of distinct editors who have edited each item. Again there is a clear a long-tail in the distribution, as there are many items that have been edited by few editors, while there are few items that have been edited by many editors. The median is 1 editor per item, while the standard deviation is 3.1 editors per item.

Finding 1.1: There is a skewed distribution of edit counts (i. e. few editors with many edits and vice versa).

Finding 1.2: There is a skewed distribution of editors per item (i. e. few items are edited by many editors and vice versa).

Figure 5.3 shows the boxplot with the counts of edits, by year in which editors started to edit. The median decreases with the years, which might be related to the fact that in the beginning there is a broader "blank space" to be edited.

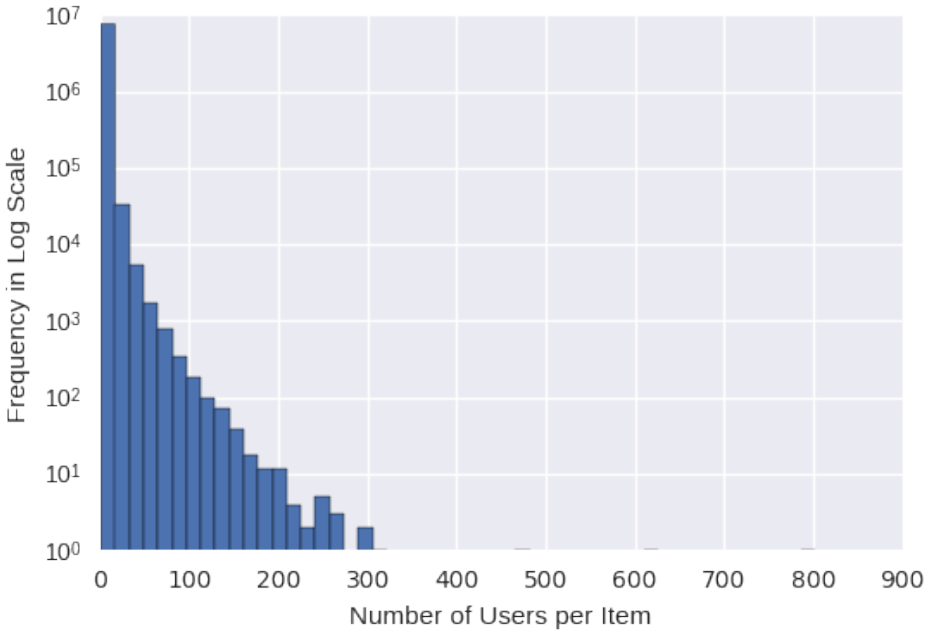

Figure 5.2: Histogram of editors per item.

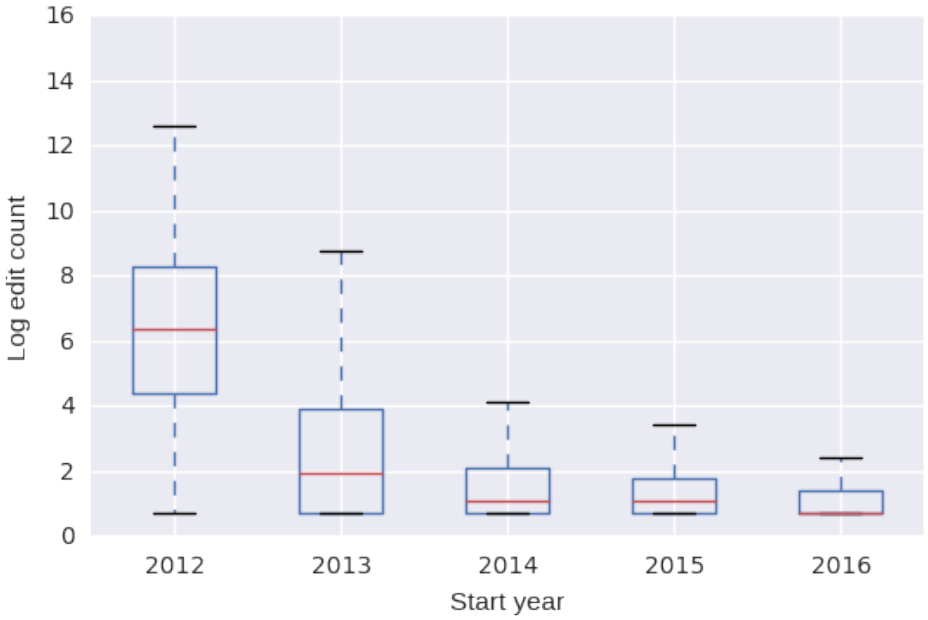

Figure 5.3: Edit counts per starting year (counts in log scale).

5.6.2 Change in the User Base over the years

When we observe the timestamp of the first and last seen edits of editors, we can analyze for each year in our dataset the number of editors who joined, as well as the number of editors who were seen last that year, and the

number of editors who joined but were seen last that year. As it can be observed from Figure 5.4, the number of people joining per year increased from 2012 until 2015, while from 2015 until 2016 this number decreased. Exactly the same behavior is observed for the last seen edit and the people whose first and last edit is seen in the same year.

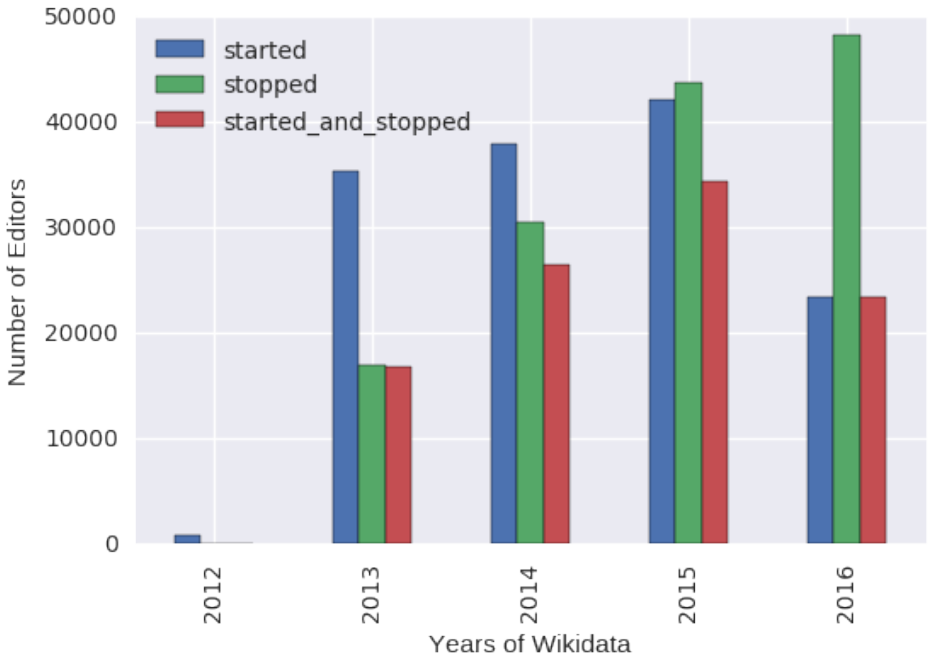

Figure 5.4: Change in User Base.

5.6.3 Editor Lifespan

Wikidata (and Wikipedia) editors often have breaks (i. e. periods of time when they do not edit). For this reason, the Wikimedia Foundation introduced the notion of *active* and *inactive* users, to differentiate between users who are present and who users who "having a pause". A user is defined as active if in the last 30 days she did 5+ edits. However, a user that is within a period of inactivity has not necessarily abandoned the project. To measure the lifespan of editors, we need to distinguish between being temporarily inactive and gone. Therefore, we look at editors as being in one of three possible status: "active", "inactive" and "gone". The distinction between active and inactive is given by Wikimedia, to distinguish those who are gone, we need an additional definition. Instead of defining an arbitrary threshold (e. g. of one year) of time after which we label editors as gone, we calculated it empirically. In order to do so, we analyzed the length (in months) of inactivity periods for all Wikidata editors. Surprisingly, the longest gap is of 16 months, which means that there were editors who, after 16 months without editing, came back to Wikidata and edited again. Looking at the percentiles we decide to use 9.967 months as a threshold in our dataset to define editors that are gone and editors that are still in the system (either in active or inactive mode). Once we labeled the dataset according to this threshold, we encountered that from the total of 140,330 editors (who edited items) 77.698 editors appear to be gone by July 2016. That is, around $55\,\%$ of the editors have abandoned the project.

To better understand the distribution of editors lifespan, we decided to analyze only editors who are gone (because only in that case we can be sure that we are looking at completed lifespans). Figures 5.5 and 5.6 shows the histogram for the editors lifespan. We can see that there are many users with a very short lifespan, and only few are long-lasting editors. There are editors who have been editing for almost 3 years.

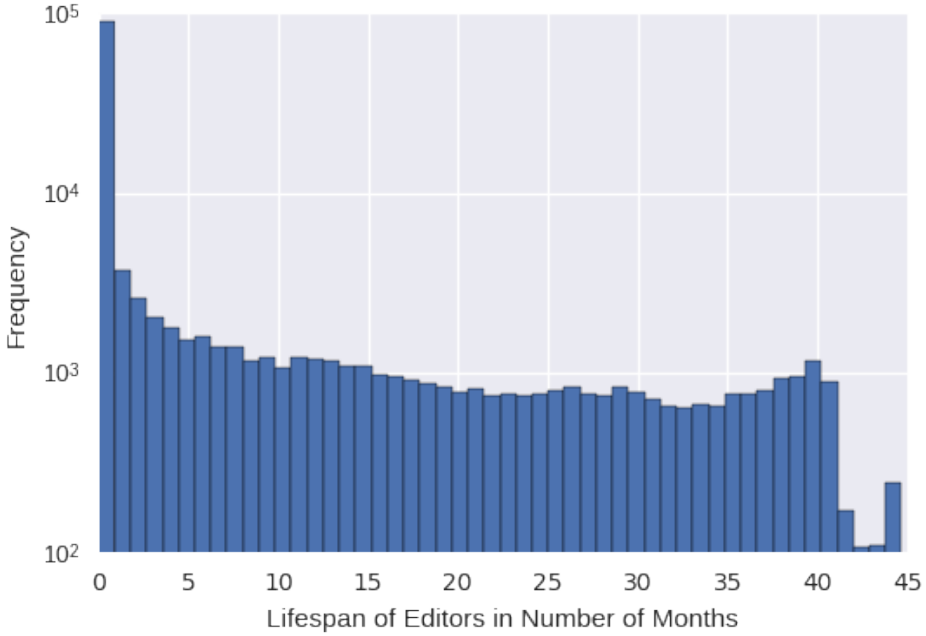

Figure 5.5: Lifespan of all Wikidata editors in our dataset.

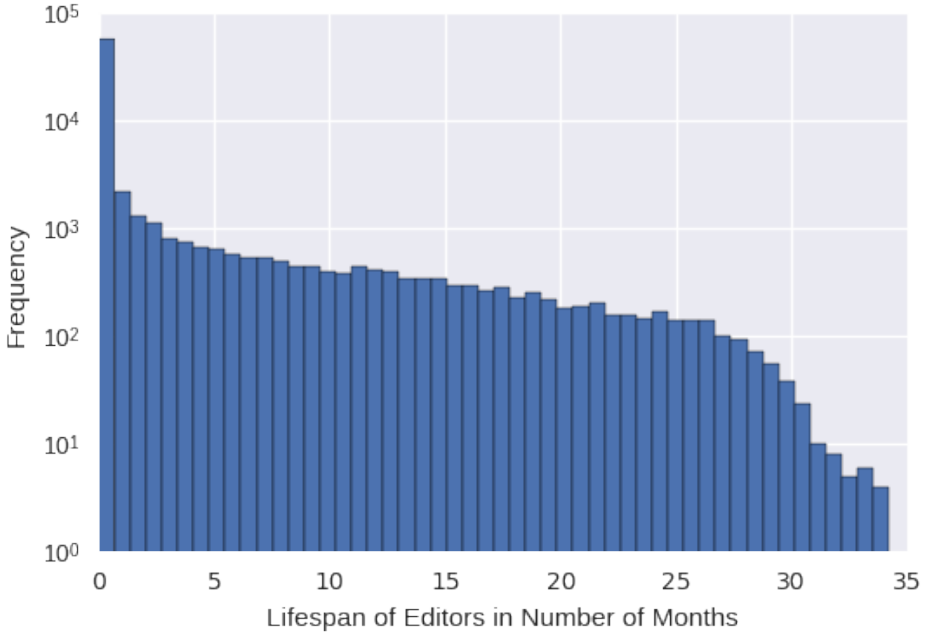

Figure 5.6: Lifespan of Wikidata editors (only users that have abandoned the platform are shown).

When comparing this drop-out ratio to other systems, we find similar results. User participation in different types of online collaborative platforms shows similar patterns to the ones we have observed in Wikidata. Generally speaking, in online communities about 90% of the participants are inactive content consumers while about only 10% actively contributes content for long periods of time (Stewart et al., 2010). Other examples include

participation in MOOCs being very skewed with numbers of participants completing the online course varying between 5% and 10% (Clow, 2013).

Figure 5.7 relates lifespan and number of edits done within the observed time frame between first and last edit. Obviously, with bigger lifespan editors may have bigger edit counts. However, interestingly the behavior is not linear, meaning that there are still people who are either slower or less committed, who have longer lifespan but the same number of edits.

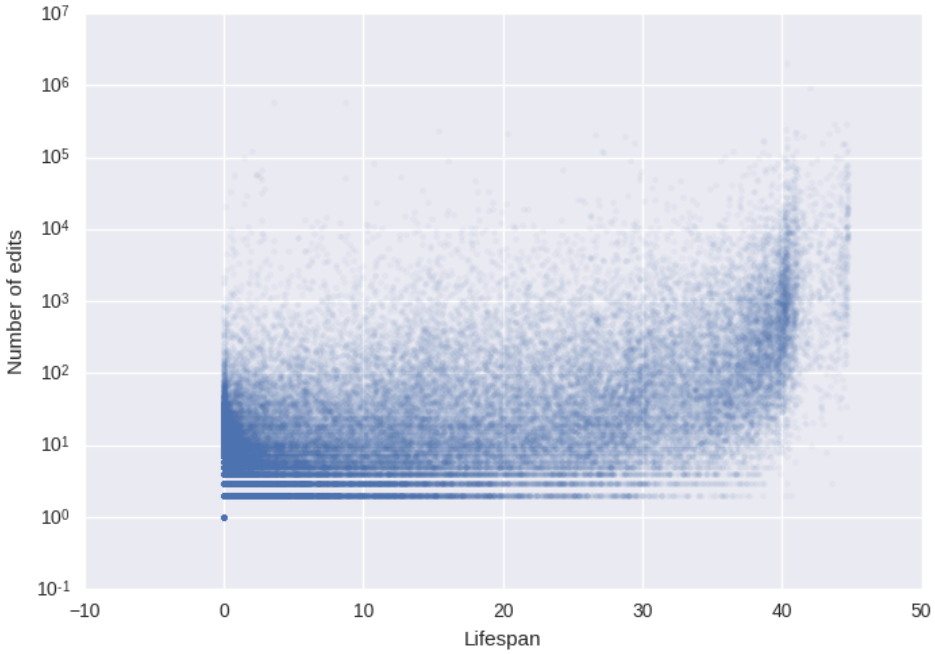

Figure 5.7: Number of edits vs lifespan.

Finding 2.1: There is a slightly skewed distribution of lifespan (i. e. few editors were in the project for many months and vice versa).

Finding 2.2: There is not a linear relation between the lifespan and the volume of edits done by editors.

5.7 Longitudinal Analysis of Wikidata Edit History

In this section we report about the exploratory analysis that we have done to understand the evolution of the editors in various dimensions. The goal of this analysis is to understand the trends that appear over time, and to find different groups of editors that should be addressed differently when designing engagement strategies.

5.7.1 Methodology

In this section, we describe the granularity of the evolution analysis, the cross-sectional indicators that we obtain for each editor over their lifetime, and the two goals that we set to distinguish between power and weak users.

5.7.1.1 Granularity of the Evolution Analysis

We analyze the evolution of editing behavior from two different perspectives: first, we look at how editors behave as they get more experienced after doing batches of edits. To group edits, we define editing sessions, a common technique in Web-based systems. Second, we observe the way editors edit over the natural months in which Wikidata has been online. Note that the latter allows us to account for the exact occurrence of inactivity periods.

Analyzing the time between edits A session is a slot of time in which a user takes a sequence of (often related) actions, that are temporally isolated from other sequence of actions. In our case, sessions are sequences of edits. In the literature we find works that analyze sessions in the context of Wikipedia (Geiger and Halfaker, 2013), but sessions have not yet been explored in Wikidata. It is interesting to do so, because the tasks that we can observe in Wikidata and in Wikipedia have a different nature and granularity. We followed the methodology defined by Geiger and Halfaker (2013), which has been widely accepted in the Wikipedia community. For each editor, we compute all the temporal differences between pairs of edits and plot a histogram (see Figure 5.8). Here we analyze all the set of registered human editors (who do not use tools), because sessions do not depend on the fact of having abandoned Wikidata or not. It is worthwhile mentioning that to generate the sessions, we consider both edits on items and edits on non-item pages (e. g. item discussion pages, user talk pages), as users tend to combine both types of edits in their sessions.

Defining the length of an editing session We analyzed the distribution of the edit differences (with logarithmic bucketing), and fitted it (with error $\chi^2 < 0.001$) as a sum of three distributions: the first peak, is a log-normal centered at around 1 second. The second distribution is an exGaussian distribution centered at around 10 seconds. The third distribution, centered at about 1 day is another log-normal distribution. We interpret the distributions analogously to what Geiger et al. did, looking at the long and short differences. From right to left, we interpret that the third distribution (with largest differences) represents the inter-session differences. The second distribution looks like the intra-session differences, whereas the first distribution with very small differences looks like intra-session edits. These small differences between intra-session edits seem very peculiar to Wikidata (and different from Wikipedia), because here users may edit multiple items simultaneously (as we will shortly illustrate). We looked at this set of edits and we identified that many of these differences referred to deletions and merge items edits. Moreover, Wikidata's GUI allows users to update multiple claims simultaneously, while giving the instruction to save all the updates after one click. Also, users may have multiple tabs open and edit interchangeably across them. The threshold for defining sessions was calculated using the aforementioned fitting distributions when considering exclusively edits done without tools and looking at the intersection of the intra- and inter-session distributions, which leads to 4.37 hours. Hence, we define a new session when two consecutive edits are separated by a time difference of at least 4.37 hours.

Finding 3.1: In Wikidata we find shorter times between edits than in Wikipedia.

Finding 3.2: We empirically define new sessions after 4.37 hours of inter-edit time, around 4 times longer than in Wikipedia.

From this model we compute the threshold that discriminates between inter-session and intra-session edit differences as about 4hours and 30 minutes. We can notice that, apart from the small second peak, the structure is similar to the Wikipedia session one, as shown in (Geiger and Halfaker, 2013). We labeled our data accordingly and grouped edits into sessions.

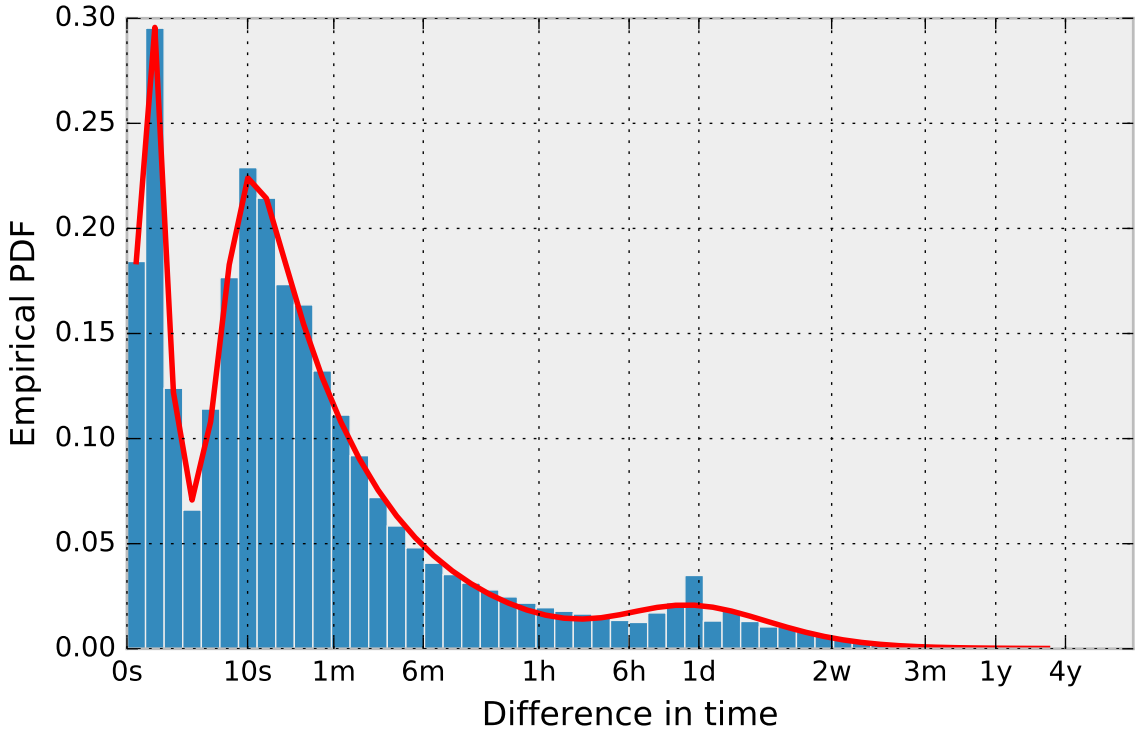

Figure 5.8: Distribution of edit differences. The red, continuous line represent the best fit with two log-normal and one expGaussian distribution.

5.7.2 Editing Behavior Indicators

We take into account the indicators used in the Wikimedia community [10] and we extend them with others defined by us. In particular, we look at three dimensions: (I) Editor contribution, (II) Editor participation, and (III) Diversity of the edits accomplished by editors.

Measuring Editor Contribution In order to measure editor contribution at each point in time we measure the following indicators:

- **i1**: number of edits. It looks at any kind of edit (from creation, to update and deletion).
- **i2**: average number of edits per item. It identifies any kind of edit done to items (either items of the schema- or the instance-level).
- **i3**: number of items edited.

Measuring Editor Participation We measure participation in terms of the time dedicated to the task of editing. The indicator that we use is:

- **i4**: number of seconds between the first and last edit in the timeframe being analysed. If there is only one edit, we decide to define i4 as 0 seconds. Note that this indicator is only relevant for a session-based analysis.

Measuring the Diversity of Tasks Accomplished by Editors To measure the diversity of of tasks that editors do, we distinguish between the different types of actions registered by Wikibase [11]:

[10] Research Metrics `https://meta.wikimedia.org/wiki/Research:Metrics`
[11] Wikibase actions `https://www.mediawiki.org/wiki/Wikibase/API`

Evolution Granularity	Editing Behaviour Indicators	Goal of Analysis
Session-based	i1, i2, i3, i4, i5	Volume of Edits
Month-based	i1, i2, i3, i5	Volume of Edits
Session-based	i1, i2, i3, i4, i5	Lifespan
Month-based	i1, i2, i3, i5	Lifespan

Table 5.2: For different options to study the evolution of editing behaviour. i1 is the number of edits. i2 is the average number of edits per item. i3 is the number of items edited. i4 is the number of seconds between the first and last edit in the session – only valid for session-based analysis. i5 is the diversity of types of edits.

- **i5**: diversity of types of edits. We use a standard measure for diversity, the the Shannon-Entropy (Shannon, 2001), to calculate the diversity of types of edits as

$$H(T) = -\sum_{t \in T} prob(T = t) \times \log prob(T = t).$$

  The probability is defined as the frequency of type t. When we compute i5, we normalize the entropy based on the number of edits done in the session/month analyzed.

Since we are interested in observing and comparing the **evolution of these indicators over time**, we first compute each of these indicators at several points in time and second, we fit a linear model using the RANSAC algorithm (Fischler and Bolles, 1981), a robust linear model estimator able to deal effectively with outliers. For each editor we obtain the values for their slope, intercept and R2. These values allow us to understand the general trend over time of an indicator for an editor. For example, when we fit a linear model on the different observations for i1, we can identify if a user is increasing, decreasing or maintaining the number of edits over time by looking at the slope of the fitted linear model. The intercept provides information about the scale and R2 indicates the error between the fitted model and the actual shape of the i1 time-based indicators.

5.7.2.1 Criteria for Identifying Power and Weak Users

The Wikidata community acknowledges the value of both (a) editors who contribute with a high number of edits, and (b) editors who contribute for a long time. As we see in Figure 5.7, these two dimensions are not necessarily always related. The optimal case is to have editors who contribute with many edits and stay a long time in the system. However, each of these dimensions separately is useful. The former implies that the knowledge base may grow or improve (depending on the concrete edits), whereas the latter means that there are editors who could eventually be available in a call for participation.

For this reason, we consider that users can be power users or weak users in two different ways: (i) in terms of the volume of their contribution and (ii) their lifespan. We set this two-fold goal as the focus of our empirical analysis, and examine the evolution of editing behaviour following the 4 different configurations listed in Table 5.2:

5.7.2.2 Empirical Findings

The main purpose of the exploratory phase of our analysis is to identify differences in the evolution of behavior between groups of editors, mainly editors with high volume of edits vs. low volume of edits, and editors with high lifespan editors vs. low lifespan editors.

Different Behaviors: Figures 5.9 to 5.15 show several scatterplots, where each depicts the slope (x-axis) and y-intercept (y-axis) for a particular indicator for each editor, in one particular evolution granularity. Figure 5.9 for

example shows the plot for indicator i1, with a session-based evolution. We decide to plot these two dimensions along the x and y axis, ignoring R2 because it does not provide a useful signal to discriminate different behaviours between these groups of editors, at least visually (in Section 5.8, we will see that even such indicator will be useful for prediction).

We observe that there are the editors with constant slope (i. e. they perform according to past sessions/months), while others have positive or negative slope, with different values for intercept (negative, zero and positive). This finding confirms that there are different evolutions of editing behaviours present among the community, where some people increase the number of edits over time, other people show a decay in the time they invest, and other editors show a constant performance.

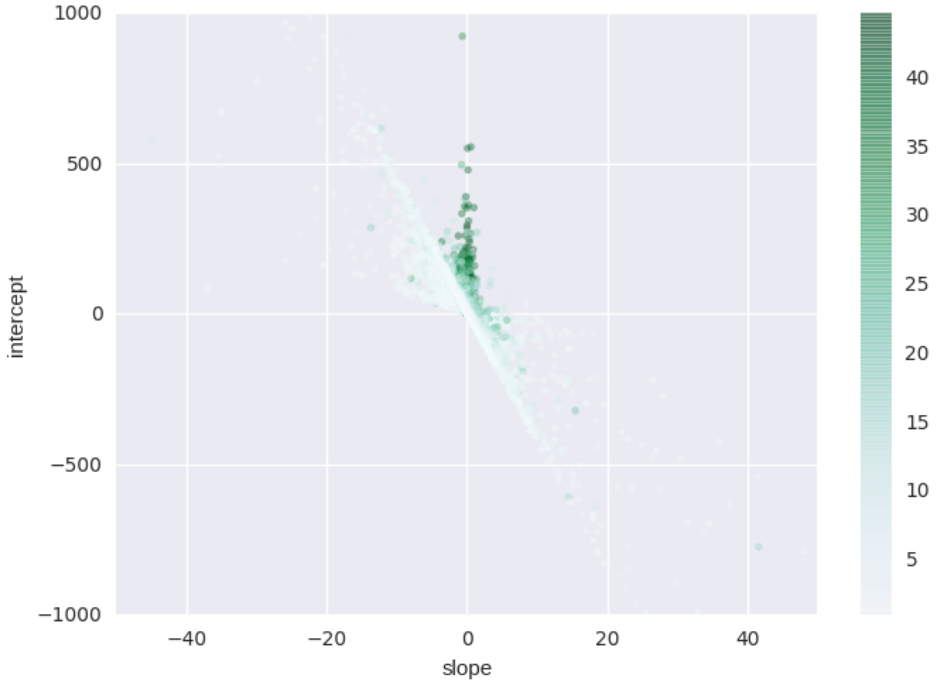

Figure 5.9: Scatter plot indicating for each editor the slope and y-intercept obtained by the RANSAC algorithm for indicator i1 (number of edits), in a month-based analysis of evolution, depicting the editors' lifespan.

The colour of each dot indicates whether the editor represented by the dot has a high or low volume of edits or long or short lifespan (depending on the particular instance). As it is clearly differentiable in Figures 5.9 and 5.10, power editors (having long lifespan), seem to have a constant behaviour in terms of the output they produce over the months (vertical green cluster in the figures). In contrast, weak editors tend to change their behaviour, having either an increasing or decreasing evolution of their contribution (diagonal white cluster in the figures). In the case of session-based evolution, most of the editors with constant contribution are editors with long lifespan or high volume of edits, but the differentiation between the two groups is not as perfect as with the monthly evolution.

Finding 4: Editors with long lifespan have a constant contribution over months, while editors with short lifespan do not.

As for the way the participation evolves over sessions, we can see in 5.12 that editors with a higher total volume of edits have a constant participation, while editors with lower volume of edits have a rather increasing

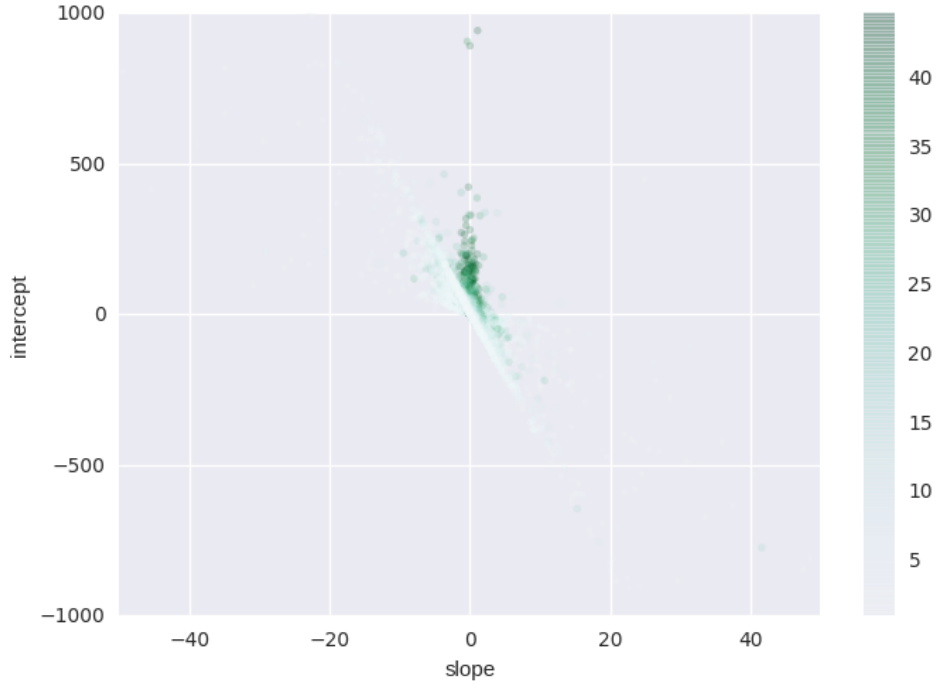

Figure 5.10: Scatter plot indicating for each editor the slope and y-intercept obtained by the RANSAC algorithm for indicator i3 (number of items), in a month-based analysis of evolution, depicting the editors' edit count.

or decreasing evolution. When it comes to comparing the participation over sessions of editors with long and short lifespan 5.13, the separation between the two groups is less clear (because some power users also have a non-constant participation evolution), but it is still visible.

Finding 5: Editors with a high volume of edits have a constant participation over sessions, while editors with low level of volume of edits do not.

Regarding the diversity of type of edits, editors with longer lifespan tend to increase the diversity over the months, while editors with smaller lifespan can either decrease or increase the diversity (cf. Figure 5.14). As for editors with higher total volume of edits, they seem to keep the diversity of the types of edits, while editors with lower volume of edits increase it or decrease it (cf. Figure 5.15).

Finding 6: Editors with a long lifespan tend to increase the diversity of the type of their edits, while editors with short lifespan can either increase or decrease it over the months.

Note that the RANSAC algorithm removes from the dataset those editors who have an insufficient number of observations to draw any conclusion on the evolution of the behavior.

Given this result, we can compare the distribution of lifespan and edit count of different slope intervals. Figures 5.16 and 5.17 show a valuable example, comparing the histogram for the lifespan of editors with a slope with an absolute value smaller or bigger than 0.2, for indicator i1 (number of edits per session). There is a clear difference in the distribution of both histograms: the editors with absolute slope value smaller than 0.2 (and therefore closer to zero), have bigger lifespans and the frequency of bigger lifespan is also higher than for the editors with absolute value bigger than 0.2.

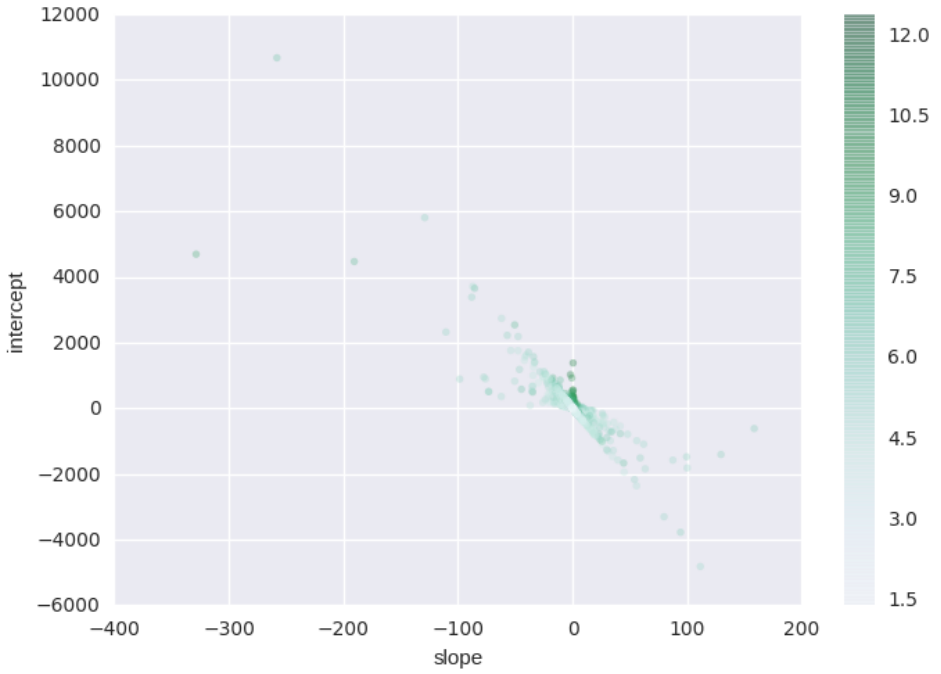

Figure 5.11: Scatter plot indicating for each editor the slope and y-intercept obtained by the RANSAC algorithm for indicator i1 (number of edits), in a month-based analysis of evolution, depicting the editors' edit count.

5.7.3 Model Building

In order to confirm our intuition about the relationship between lifespan/edit count and the indicators, we use a Random Forest regressor with the whole dataset as training data and visualize the predicted values on the indicator space.

In Figure 5.18 and 5.19, we visualize a projection on the slope and intercept of the predicted lifespan for i1 and i2 respectively of the month based indicators. The predicted values corroborate the intuition about the absolute value of the slope as being a good indicator of high lifespan: for i1, a slope close to zero (constant behaviour) corresponds to high lifespan; for i2, we see that the second and fourth quadrant are associated with low lifespan, following the pattern visualized in Figures 5.9- 5.15. The corresponding projection for the session based indicators (Figures 5.20 and 5.21) are less clear, presumably because the interaction between all different indicators are more involved, and a two-dimensional projection is not able to simply visualize them. We verify the prediction capabilities of such model in the next section.

5.7.4 Hypotheses Revision

As we can see along the Figures of Section 5.7.2.2, the difference between editors with long / short lifespan and editors with high / low volume of edits, in terms of the way they evolve is in some cases obvious and in others not. We can confirm Hypothesis 1, because we see that editors with high lifespan show a constant contribution, while other do not. The same applies to high volume of edits, although with less strength (cf. Finding 4). We can also confirm Hypothesis 2 in the case of lifespan, because we see that people with longer lifespan maintain a constant participation, while editors with shorter lifespan do not. However, when it comes to volume of edits, the measurements do not help differentiating the two groups of editors as clearly. Hypothesis 3 is rejected, because we see that both in long and short lifespan, and in high and low volume, editors tend to increase the diversity of the type of actions they accomplish. Hence, evolution in contribution and evolution in participation are good

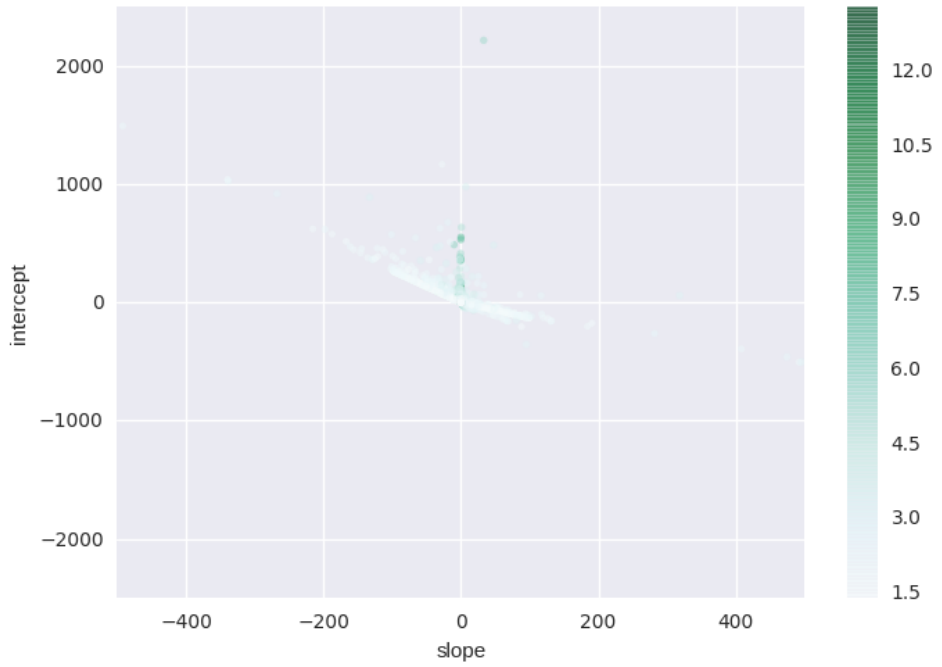

Figure 5.12: Scatter plot indicating for each editor the slope and y-intercept obtained by the RANSAC algorithm for indicator i4 (seconds per session), in a session-based analysis of evolution, depicting the editors' edit count.

indicators to differentiate standard and power editors in (especially in terms of lifespan), while diversity alone it is not sufficient.

5.8 Predicting Volume of Edits and Lifespan of Editors

Motivated by the findings presented in the previous section, we define our prediction tasks as follows: Given an editor, and a set of observations based on the aforementioned productivity, participation and diversity indicators, we predict

- whether the editor will be contributing with a high or low volume of edits and
- whether the editor will contribute for a long or short lifespan.

To solve these two prediction problems, we use supervised learning methods. More specifically, we use binary classifiers that take as input the information about the slope, intercept and R2 obtained by applying the RANSAC algorithm on the multiple indicators-based measurements, across sessions and across months. To define the thresholds for high / low volume of edits and long / short lifespan, and create the 4 classes (i. e. power and standard editors as for lifespan, and power and standard editors as for volume of edits), we observe the distribution of both volume of edits and lifespan, and decide that 15 months and 100 edits are suitable numbers to empirically distinguish between the different classes of editors that we define in terms of volume of edits and lifespan respectively.

We select two different classifiers: Random Forest[12] and a Logistic Classifier to compare their performance in terms of precision, recall, f1-score, and support. We evaluate the classifiers in different settings: (a) for a session-based evolution, we run the prediction evaluation with training data of size 100, 200, and 300 sessions. (b) For

[12]The Random Forest parameters chosen are: 100 estimators and bootstrap technique with subsample class balancing.

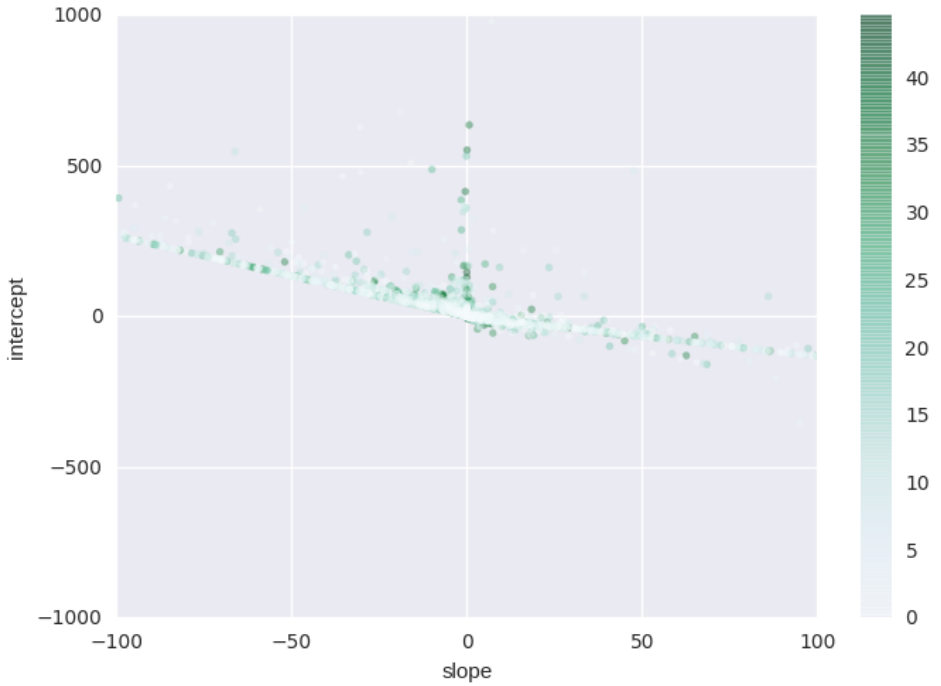

Figure 5.13: Scatter plot indicating for each editor the slope and y-intercept obtained by the RANSAC algorithm for indicator i4 (seconds per session), in a session-based analysis of evolution, depicting the editors' lifespan.

a month-based evolution, we pick training data of size 3, 5, and 10 months. The data that we use at this point is the complete set of edits done by humans without tools over items. We considered filtering out editors whose status is not 'gone', as in such case we cannot really state their lifespan with certainty. However, we noticed that after applying the RANSAC algorithm, there were no editors whose state is active and her lifespan is shorter than 15 months. The only editors who are not gone, have a lifespan longer than 15 months and therefore can also be labeled as having a long lifespan (no matter what the exact final value for the lifespan will be).

Figures 5.22 and 5.23 show the average F1-scores obtained for each class (power and standard) by the two classifiers, after running 10-fold subsample validation, for each training data size. In all cases, the Random Forest classifier outperforms the Logistic Classifier. If we compare the two plots of Figure 5.22 to the two plots of Figure 5.23, we observe that the former show stable F1-scores, even when augmenting the number of months or sessions in the training data. In the latter plots, there is a trend to increase the F1 with a bigger training data size.

According to the results, we obtain a higher F1 when we predict lifespan than we predict volume of edits.

5.9 Discussion

In this section we discuss the findings highlighted in sections 5.6 and 5.7, and present some of the implications of our research results for the Wikidata community.

5.9.1 Summary of Findings

In summary, with this work we found that:

- There is a skewed distribution of edit counts.

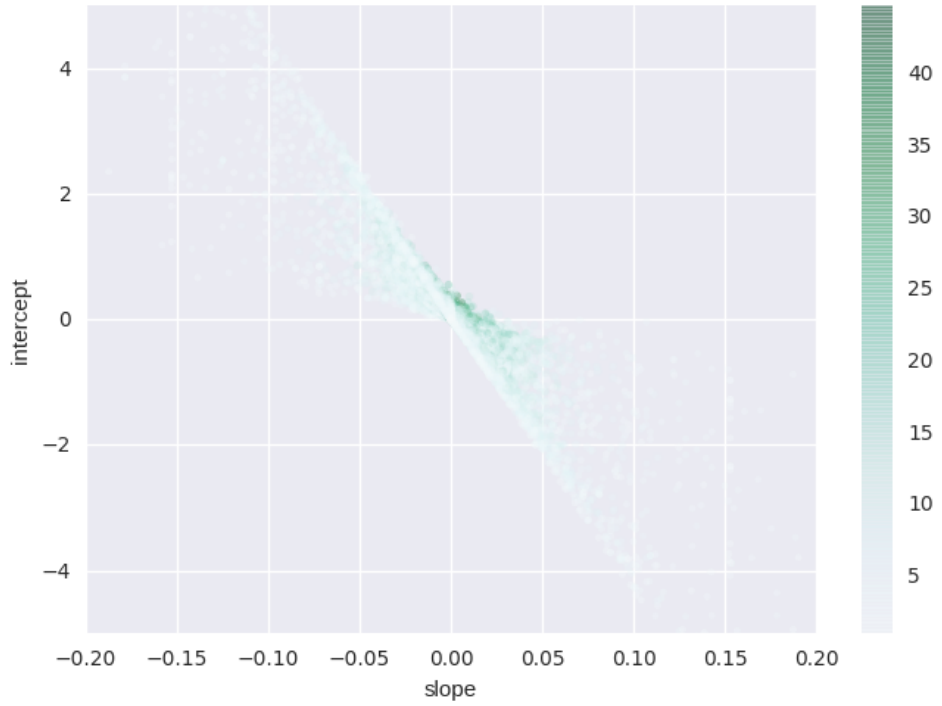

Figure 5.14: Scatter plot indicating for each editor the slope and y-intercept obtained by the RANSAC algorithm for indicator i5 (diversity of type of edits), in a month-based analysis of evolution, depicting the editors' lifespan.

- There is a skewed distribution of editors per item.
- There is a slightly skewed distribution of lifespan.
- There is not a linear relationship between the lifespan and the volume of edits done by editors.
- In Wikidata we find shorter times between edits than in Wikipedia.
- We empirically define new sessions after 4.37 hours of inter-edit time, around 4 times longer than in Wikipedia.
- Editors with long lifespan have a constant contribution over months, while editors with short lifespan do not.
- Editors with a high volume of edits have a constant participation over sessions, while editors with low level of volume of edits do not.
- Editors with a long lifespan tend to increase the diversity of type of their edits, while editors with short lifespan can either increase or decrease it over the months.

5.9.2 Interpretation of Findings

Participation inequality (Yasseri et al., 2012) is present in a vast amount of Web systems, where often only the 1% of users contribute heavily, 9% of users contribute sporadically and the remaining 90% are so-called lurkers, who consume information (reading and observing) but do not actively contribute. Hence, the skewed distribution of edit counts is an expected behavior. Still, it is a relevant descriptive statistic that can be used to understand the order of magnitude of the set of editors whom should be addressed, and define parameters of the editor retention

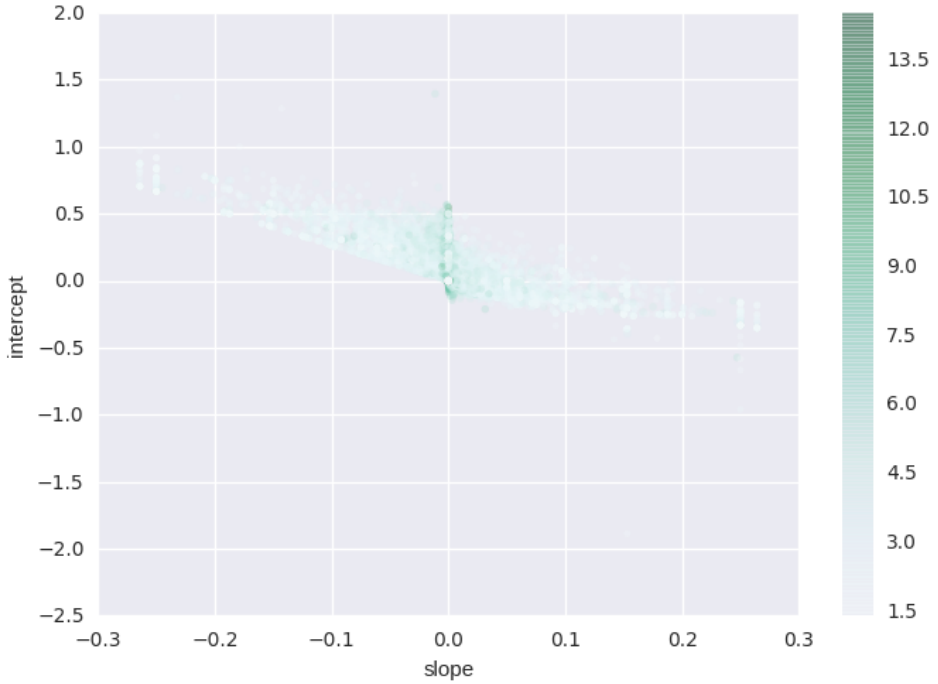

Figure 5.15: Scatter plot indicating for each editor the slope and y-intercept obtained by the RANSAC algorithm for indicator i5 (diversity of type of edits), in a month-based analysis of evolution, depicting the editors' editcount.

/ reactivation strategy (e. g. an upper bound for the number of expected edits to be improved). The (also) skewed distribution of editors per items indicates that not many item descriptions are really crowdsourced, which in the case of Wikidata might be especially detrimental, considering that the knowledge base is meant to be the aggregate of what primary Web sources state. Features like qualifiers, ranks and references in statements allow Wikidata to portray existing plurality, and therefore, the more people involved in curating information about an item, the higher the chance to capture this plurality and the lower the risk of having certain kind of bias during the data collection.

The distribution of the lifespan shows that there are a couple of thousands of people who have been contributing for almost the 4 years that we studied, which is priceless. Again, having this descriptive statistic can help configure the retention / reactivation strategy. Currently, Wikimedia encourages editors after they reached editing milestones (e. g. editors get congratulated via the wiki after they achieve their 100th edit). So, similarly, we could think of acknowledging people for being in the project for a particular amount of time. The fact that the relationship between lifespan and volume of edits is not linear indicates that there are some people who make a contribution of for example a thousand of edits in a short time – even less than one month. Probably events such as hackathons and editathons with a specific focus (e. g. to enter the description of women Swiss scientists in Wikidata) stimulate editing activity of participants, but in some cases it might mostly during the event. Moreover, data providers might also edit in bulk for a short time to ingest one single dataset into Wikidata.

Having shorter times between edits than in Wikipedia is coherent with the nature of Wikidata, because it allows people to edit structured data – that being adding references, updating a date or a quantity, or creating a link between two items. In Wikipedia, people can also correct a typo, add a citation or a link, but the main task is to write a text (either a complete page or a paragraph), and the edit is registered when the text is saved, and not after each word has been edited. In Wikidata, if statements are edited within the same item, and the editor is proficient it is very much feasible that two edits are accomplished within one or few seconds. The high number of consecutive edits done in or under 5 seconds can be due to the fact that some actions, like merging

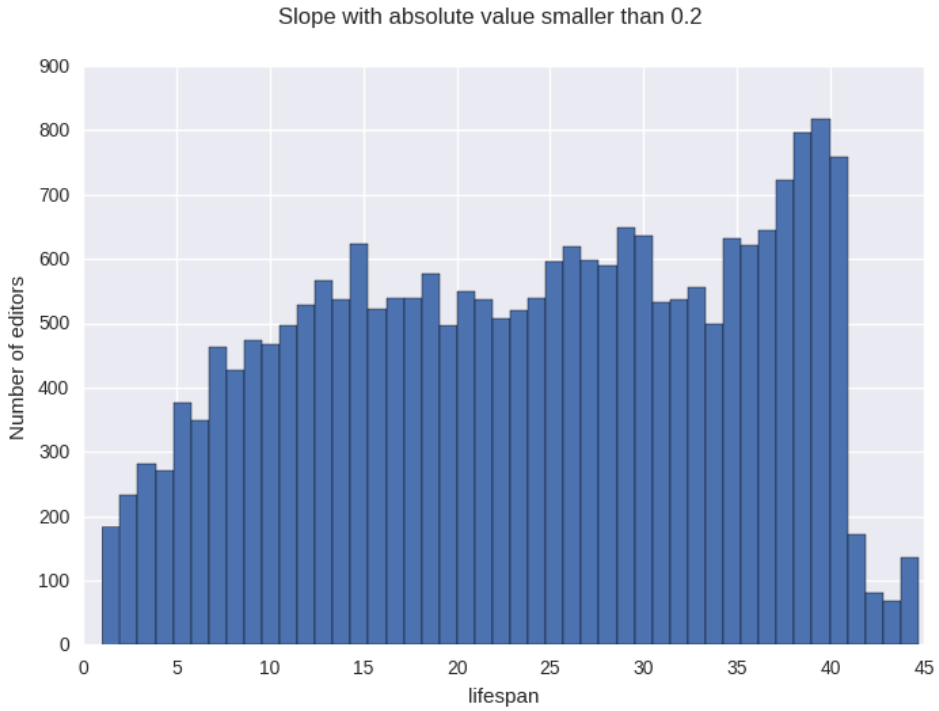

Figure 5.16: Histogram showing the lifespan for editors with slope of an absolute value smaller than 0.2.

items and deleting items (only granted to special editors) can be executed one after the other in a very short time. Furthermore, labels in various languages can be edited all at once. It can also be that some editors edit multiple items in parallel (e. g. in several tabs open at the same time) to perform similar kind of actions, like for example geolocate, or add descriptions. Another explanation for the high number of very short times between edits is that we are able to identify edits by tools that leave a register trace (e. g. # petscan), but it is not possible to filter out API calls executed from tools that other developers might have implemented and did not tag or even advertise.

The results referring to the differences between power and standard editors suggest that power editors have habits in terms of their contributions – edits done each month – and the participation – the time spent in sessions. The existence of habits reflects not only the conviction to contribute, but it also shows that these editors successfully manage to find work they can do. From conversations we had with editors, we know that some power editors add information about the film they have watched, or edit information about the person they heard about in the news. Others look everyday through the list of recent and unpatrolled changes. There are different events that trigger editing action, and power editors follow them regularly. One could argue that if contribution and participation are both constant, then there is no clear signal of a learning effect that leads to editors becoming faster in accomplishing their tasks. We assume that the explanation is that power editors have a learning effect mostly in the beginning and later their efficiency becomes stable. Since they typically have many edits, it is possible that this initial learning effect is invisible to the linear model fitted. The increasing trend for the diversity of the types of edits present in power editors is aligned with the fact that intrinsically motivated people tend to "seek out novelty and challenges, to extend and exercise one's capacities, to explore, and to learn."(Ryan and Deci, 2000).

Between the two prediction problems that we set up (i. e. predicting the lifespan range and predicting the range of volume of edits), predicting the lifespan has a higher priority, because it gives us the key information about when we should address standard editors that will become inactive. Therefore, being able to predict lifespan more accurately than the volume of edits is a positive result.

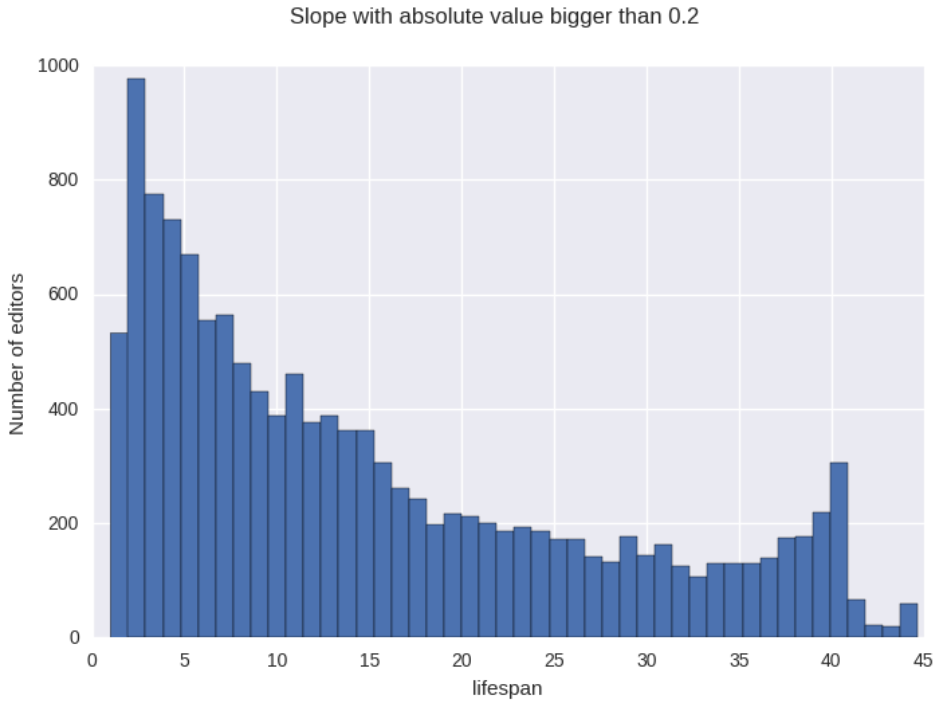

Figure 5.17: Histogram showing the lifespan for editors with slope of an absolute value bigger than 0.2.

5.9.3 Implications

Based on our findings, we identify three major areas where the Wikidata community could focus:

- Increasing the crowdsourcing ratio per item: in the same way that there are indicators to draw attention to the relative (in)completeness of items (see the ReCoin tool `https://www.wikidata.org/wiki/User:Ls1g/Recoin`), there could be indicators measuring the plurality of item descriptions and the number of editors involved in the description (as a crowdsourcing ratio). The goal of showing such information would be to request the help of further editors and improve the plurality of the item description – however that is defined.

- Acknowledging long lifespan and high activity periods: the retention literature conveys that it is better to interact with users before they leave because the probability of making them go back to the system once they drop-out are lower. Hence, we encourage the Wikidata community to add lifespan and high activity recognition to the acknowledgments that are implemented in the wiki. The system could congratulate editors for being in the system for a long time or for having a high activity peak. The message could report historic statistics or even show flashbacks about a significative edit done by the editor a while ago. As a reward, the community could grant these editors a special privilege in their "wikidata-versary".

- Encouraging a behavioral change that makes standard editors become power editors: with Wikidata's increasing size and complexity, it is becoming more and more challenging for editors to master the variety of tools to edit, query and visualize Wikidata's data, and find things they can contribute to. In our work, we found out that a very high rate of editors have a very short lifespan, and their evolution show that their contribution and participation is not constant. That is, there are many editors with the potential to become more active. Usually, early dropouts are motivated by any of the multiple possible phenomena such as people lacking the conviction for free knowledge, people not finding the way to contribute and

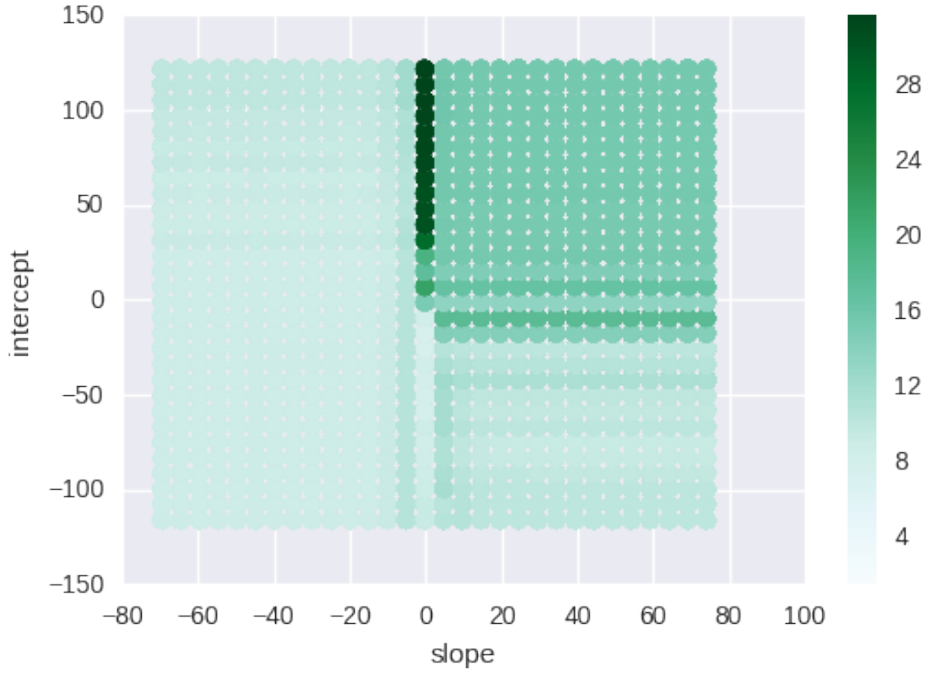

Figure 5.18: Random Forest lifespan prediction on a projection of slope and y-intercept of i1 for the month-based indicator.

Wikidata not being able to develop a sense of "addiction"[13]. Organizing training and dissemination events to educate people about the value of free knowledge is the solution to (i) and that is something that Wikimedia already does. For (ii) and (iii), Wikidata needs to provide a methodology and tools that allow these editors find valuable work to do and develop a reinforcing editing habit. Previous works have shown that suggesting things to edit in Wikipedia can be useful both for the users and the system (Cosley et al., 2007; Wulczyn et al., 2016), and providing article feedback can also help readers transition into editors (Halfaker et al., 2013). Given that the dynamics defined in the Wikipedia community is similar to the Wikidata community, we believe that there is space to develop a solution that can guide Wikidata editors and help them become more active and transition close to power editors (if they are willing to do so). We propose to design a system that helps standard editors find their editing mission. Missions could be defined individually or collectively (i. e. shared with other editors). To encourage a change in their behaviour, it would be important to consider behavioral change theories (Yasseri et al., 2012; Michie et al., 2011) suggest that change is more effective when the person frames intentions and goals, has the chance to self-regulate him/herself, and can freely select from available choices. So, as a design principle, the system would let them define what they would like to achieve, and decide what they finally would like to work on let, rather than impose or assign work to do. The system, still, would need to reduce the number of editing possibilities to a manageable and attractive set of options. And such an algorithm would need to exploit the main difference between Wikidata and Wikidata: the structure in the data and the fine granularity of traceable actions. Another key feature of such a solution could be a tight collaboration between power and standard users. Standard editors would define their intentions (e. g. editing for the city of Zurich) and identify themselves with roles (e. g. someone would like to become a quality ninja, but she does not know how yet). Power editors would set calls for actions and define data needs. The system could enable a 1:1 contact between power and standard editors, so as to share and disseminate

[13]Causes to drop out in Wikipedia by the community `https://www.wikizero.com/en/Wikipedia:WikiProject_Editor_Retention#`

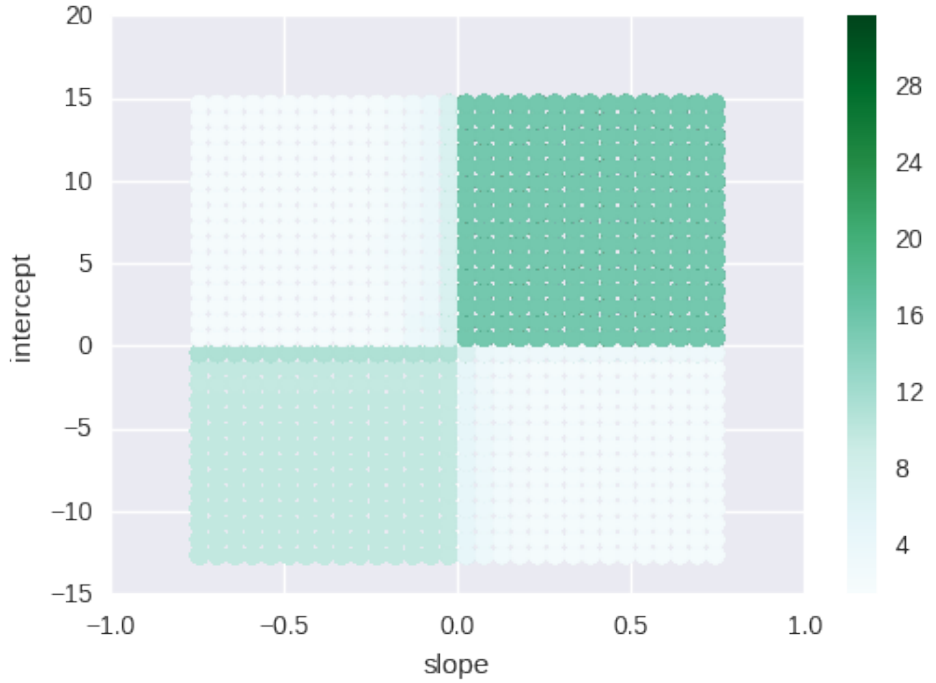

Figure 5.19: Random Forest lifespan prediction on a projection of slope and y-intercept of i2 for the month-based indicator.

best practices, recommendations on habits and know-how. Wikimedia currently enables mentorship [14]. Our system would aim at systematizing some actions in this process. The predictive models defined in this paper would be useful to estimate the amount interaction (and help) that editors would need, as well as to compute the time when the system interacts with the editor, whose behaviour we would like to improve (in terms of engagement). [15]

5.9.4 Limitations

Our work has two major limitations: *First*, we do not provide a qualitative analysis of the edits. While we distinguish among different types of edits (see indicator i5 5.7.2), we focused on contribution, participation and diversity of the edits without observing features of the items (e. g. topics and categories of the items), or the quality of the actual edits (e. g. whether the new statement created by an editor is semantically accurate). Clustering items by topics would be useful for computing further diversity indicators, whereas understanding the quality of the edits would help us categorize editors in different ways. Labeling the quality of edits can be beneficial for filtering out editors whom we may not want to retain – people who intentionally provide incorrect, hence, disruptive edits – and consequently can help us design more accurate measures against attrition. Yet, predicting the lifespan is a useful information by itself, because it gives a hint about the moment when we need to intervene to encourage behavioural change. Likewise, the prediction of the volume of edits gives an indication about the magnitude of the editor's work. The fact of knowing if they are primarily good or bad edits will only change the interpretation and the way we will proceed (e. g. in the case of malicious editors, the shorter the lifespan and the lower the volume of edits, the better; and in the case of helpful editors it is exactly the opposite)

[14] Wikipedia Mentorship `https://en.wikipedia.org/wiki/Wikipedia:Mentorship`

[15] This idea, together with the major findings of this research were presented in a talk at WikidataCon `https://goo.gl/vKH1kj`. The Wikidata community appreciated the findings and welcomed this proposal to improve editor attrition.

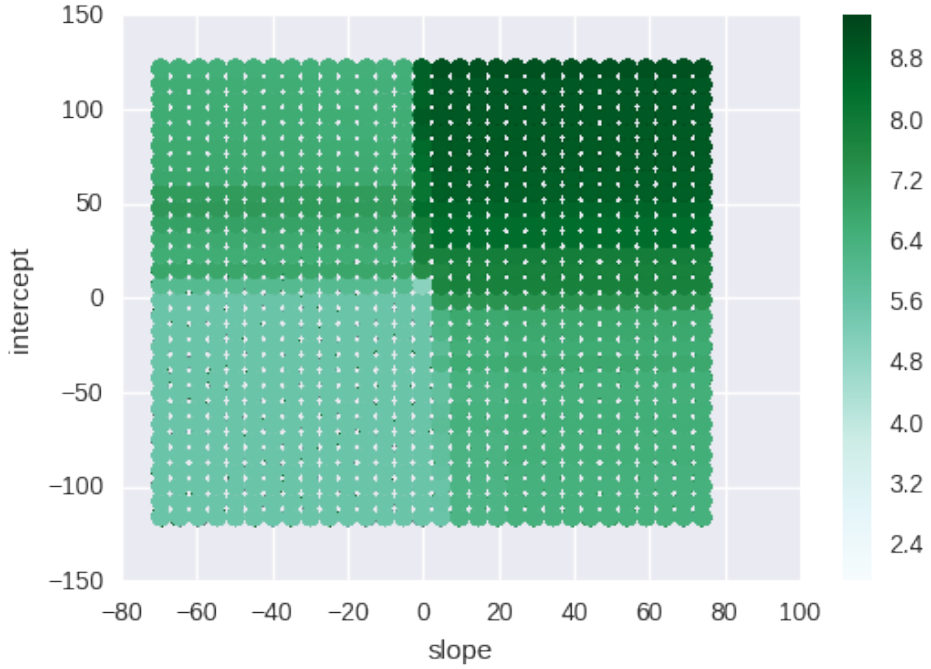

Figure 5.20: Random Forest count edits (log scale) prediction on a projection of slope and y-intercept of i1 for the session-based indicator.

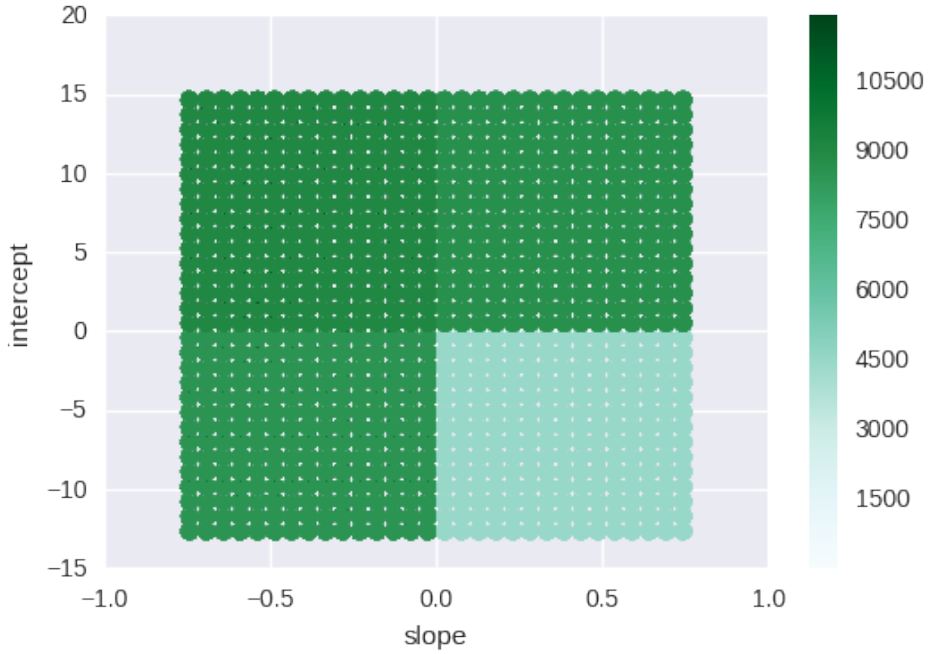

Figure 5.21: Random Forest count edits prediction on a projection of slope and y-intercept of i2 for the session-based indicator.

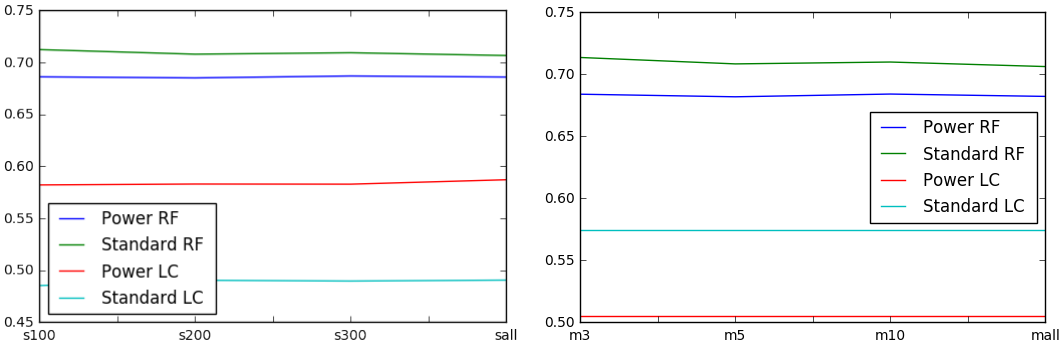

Figure 5.22: Plots comparing the F1-score for each class (power and standard) obtained by two classifiers (a Logistic Classifier and the Random Forest classifier) when predicting the volume of edits that editors will make. The first plot shows the F1-score evaluation using 100, 200, 300 sessions of edit history per editor as training data, while the second plot shows the evaluation using 3, 5, 10 months of edits per editor as training data.

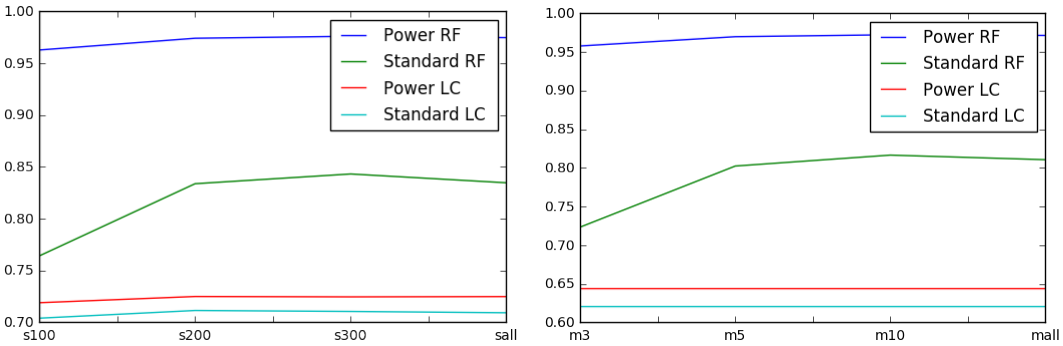

Figure 5.23: Plot comparing the F1-score for each class (power and standard) obtained by two classifiers (a Logistic Classifier and the Random Forest classifier) when predicting the lifespan that editors will have, using 100, 200, 300 sessions (first plot) and 3, 5, 10 months (second plot) of edit history per editor as training data.

Second: we do not reveal the reasons that lead to editor attrition. Understanding the issues that standard editors face is crucial for designing a solution to the attrition problem. That is why, we plan to run a survey in the Wikidata community, to address this question.

While these are natural extensions of our work, they are complex topics that are worth standalone articles. Automatically identifying the quality of edits, for example, is a highly challenging task (Sarabadani et al., 2017). In fact, the definition of data quality in Wikidata is still under debate among the research and volunteer communities [16]. Similarly, partitioning the knowledge base according to different topical domains could be done in many various ways and would require studying the application of specific techniques such as topic modelling in knowledge graphs. These two tasks are out of the scope of this study, which by design aims at understanding the differences in participation and dedication between power and standard editors, and predicting the group to

[16] `https://www.wikidata.org/wiki/Wikidata:Requests_for_comment/Data_quality_framework_for_Wikidata`

which editors will belong in the future.

5.10 Conclusions and Future Work

In this paper we have performed a longitudinal cross-sectional analysis of Wikidata's editor community behaviors aiming at understanding how different types of editing behaviors lead or not to high volume and long-term engagement with the community.

Our results show that:

- The number of new editors joining the Wikidata community has been increasing over time, but it decreased recently.
- The distribution of contributions is very skewed with few editors contributing most of the edits and many editors performing just a few edits.
- To correctly define an edit session, it is necessary to consider the intra-session edits distribution (as well as intra-session differences and inter-session differences), obtaining a threshold of 4.3 hours, around 4 times longer than in Wikipedia.
- There is no linear relation between editors' lifespan and the volume of edits they provide.
- Power editors tend to show a constant contribution over the months and a constant participation over the sessions, while standard editors show instead an increasing or decreasing tendency.
- Power editors tend to increase the diversity of the types of edits, but this dimension alone is not clearly separating the two sets of editors, because some standard editors also increase their diversity over time.
- Despite the unbalanced nature of the data (i. e. few editors with many edits or long lifespan), it is possible to automatically predict the future the volume of edits and lifespan duration of an editor based on the available edit history of Wikidata editors. We are able to obtain better prediction results than a naive classifier in each of the 4 tested configurations. We are able to predict lifespan better than volume of edits (with an average F1 score above 0.9) in both session- and month-based evolution.

Our results are relevant to the Wikidata community, because they shed some light on the way power and standard editors in Wikidata evolve differently over time. Having these insights is useful –especially now that there is still limited knowledge about the way the community progresses– to design methods that encourage standard editors to contribute more, and hopefully also longer, as they experience progress. Additionally, our results and observations may be of use to other crowdsourced knowledge curation and maintenance projects with similar characteristics to better engage their communities of contributors.

While this work is leveraging a very large dataset of activity logs, in the future we plan to complement our work with a qualitative study performed by surveying and interviewing representative Wikidata editors sampled from the different categories we looked at in this work. As future work, we can also include the *quality of edits* as a feature which can be measured with the revert rate as a proxy. Furthermore, as a follow-up work, we plan to work on the implementation of the solution for attrition and retention management proposed in Section 5.9, in collaboration with Wikimedia.

5.11 Acknowledgments

We would like to thank Michele Catasta for his feedback at an early stage of this research, and the rest of the participants of our Dagstuhl Research Meeting "Crowdsourcing Research - Transcending Disciplinary Boundaries". We also would like to thank Michael Luggen for his help to set up one of the machines used for the experiments of this project. This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 732328, as well as from the COST Action IC1302 - Keystone.

Chapter 6

Crowd Work CV: Recognition for Micro Work

Abstract: With an increasing micro-labor supply and a larger available workforce, new microtask platforms have emerged providing an extensive list of marketplaces where microtasks are offered by requesters and completed by crowd workers. The current microtask crowdsourcing infrastructure does not offer the possibility to be recognised for already accomplished and offered work in different microtask platforms. This lack of information leads to uninformed decisions in selection processes, which have been acknowledged as a promising way to improve the quality of crowd work. To overcome this limitation, we propose Crowd Work CV, an RDF-based data model that, similarly to traditional Curriculum Vitae, captures crowd workers' interests, qualifications and work history, as well as requesters' information. Crowd Work CV enables the representation of crowdsourcing agents' identities and promotes their work experience across the different microtask marketplaces.

6.1 Introduction

One of the challenges in human computation systems is to involve the humans who, as intelligent and independent beings with a particular knowledge, are crucial to solve problems that machines can hardly solve alone. Crowdsourcing alleviates this challenge, as it provides a mechanism to distribute a task among a potentially large group of people who subscribe to an open call on the Web (Quinn and Bederson, 2011). A promising strategy to improve the quality of crowd work, which is particularly relevant for knowledge-intensive crowdsourced tasks, is to find the most suitable worker(s) for a microtask (or vice versa), as Kittur et al. (2013) highlighted. With the current increasing order of magnitude (in available micro work and workforce), and the evidence of the crowd being diverse in terms of background (Ross et al., 2010), personality (Kazai et al., 2011), and motivation (Kaufmann and Schulze, 2011), analysing different aspects of the agents involved in crowdsourcing in order to improve microtask assignment accordingly becomes meaningful. However, the realisation of such process is hindered by the current microtask crowdsourcing infrastructure, which is highly focused on independent marketplaces. Even if many of them have adopted some common patterns (e. g. consider majority voting as aggregation method), each of them acts as a data silo. When crowd workers are registered and work in several marketplaces, the work they perform is registered in the marketplace they worked at and only visible there. If a requester is interested in knowing further information on the achievements and proven skills of the worker (e. g. through obtained qualifications) in other marketplaces, this information is not accessible programmatically, even though the data exists and it is visible to the worker. The same applies to requester information. This lack of data interoperability between marketplaces has a negative impact in the process of finding the best combination of workers and microtasks and may result in uninformed decisions. In Section 6.2 we describe a motivational scenario in more detail.

To overcome this limitation we propose Crowd Work CV, an RDF-based data model to represent someone's crowd work life, equivalently to what traditional Curriculum Vitae reflect. Crowd Work CV enables the aggrega-

tion of valuable information about crowd workers and requesters, which may be exchangeable if the data owner—the agent represented in the Crowd Work CV—decides to do it. The approach is conceived to boost transparency among the crowdsourcing agents, which has a positive effect in crowdsourcing environments, too (Huang and Fu, 2013). The contributions of this work are:

- The definition of a conceptual model to represent Crowd Work Curriculum Vitae information (see Section 6.3)
- The implementation of the data model into an OWL vocabulary (see Section 6.3)

The definition of the Crowd Work CV data management system is out of the scope of this paper.

6.2 Motivational scenario

Let us imagine Alice, who has registered in several marketplaces[1] and her experience is being assessed for a microtask for labelling the sentiment of Spanish Web sites, because the requester would like to trust the responses of experienced crowd workers more than these from unexperienced crowd workers. **1)** Alice registered at **ClickSense** but her work history is still empty. **2)** Alice has been working on text translation at **Neobux**, where she has obtained a Spanish qualification defined by a requester. **3)** At **GetPaid** Alice successfully completed several microtasks that actually belong to the same group of microtasks where she is going to be assessed, because when publishing the whole group of microtasks, CrowdFlower distributed some of them to ClickSense and some to Neobux. **4)** At **MTurk** Alice worked with very good performance on microtasks whose purpose was to analyse the sentiment of Tweets). Even if the type of content analysed is different the task of deducing the sentiment of text is equivalent.

Alice will be poorly assessed within the context of ClickSense—where the requester is evaluating the experience of workers—because she did not work there before. Other crowd workers to compare with have worked on Web site sentiment analysis microtasks, but with a much lower performance than what Alice did at GetPaid. She has a language qualification, and she has proven to be capable of solving the type of job being analysed, and even other related microtasks dealing with a similar problem. Still, due to a lack of shared information, the requester will not consider her work experience. This affects negatively to both sides: the requester is not taking advantage of a potentially good worker for the task at hand, and the crowd worker is missing an opportunity to work on something she might be interested in because of its similarity to previous completed crowd work.

6.3 Modelling the Crowd Work CV

With Crowd Work CV, we aim at adopting the procedure of reporting work experience from the traditional workplace, where there are plenty of guidelines about the information that should be included in CVs. We identify 5 requirements for a CV in microtask crowdsourcing: First, **domain independence**, because to enable the reusability of crowd work activity reports, a clear separation between the domain knowledge and the management of crowd work history needs to be ensured. Second, **marketplace independence**; the model needs to guarantee a certain level of generality, representing well-established processes instead of particular isolated characteristics provided by one particular marketplace. Third, **semantic and syntactic interoperability**; an agreement on vocabulary should be ensured using a common (in our case) Web-based syntax. The semantics should be explicitly defined and shared separately from the data. Fourth, **extensibility**, because the appearance of new features in marketplaces, or the definition of new workflows in crowd work should not interfere in the already specified model and existing crowd work CV descriptions. Fifth, **compatibility with traditional CV information** defined in standard systems like Europass and LinkedIn[2].

[1] GetPaid `http://www.getpaid.com`, Neobux `http://www.neobux.com`, MTurk `http://www.neobux.com`

[2] Europass `http://europass.cedefop.europa.eu/en/documents/curriculum-vitae` and LinkedIn `https://www.linkedin.com/`

6.3.1 The Crowd Work CV ontology

The Crowd Work CV ontology describes crowdsourcing agents (i. e. crowd workers and requesters), their **interests**, obtained **qualifications** and **work history**. The ontology is available online, and written in OWL[3]. We followed the ontology engineering methodology proposed by Noy and McGuinness (2001) and considered reusing related ontologies. We decided to reuse some classes and properties of FOAF[4] for the description of agents and SIOC[5] for the description of user accounts, because their definition fits directly our needs and information annotated with such vocabularies on the Web becomes reusable. While the Crowd Work CV elements share some commonalities with the ResumeRDF[6] ontology, for modularity reasons, we decided to define our own ontology elements (which are more oriented to crowdsourcing) and align the CV concept to this ontology. We list the most relevant elements in the Crowd Work CV ontology and describe their purpose. Figure 6.1 shows the graphical representation of the ontology.

CV is the core class of the ontology. It aggregates all the information that is used to report the crowd work life of an `Agent`, which (from FOAF) can be either a `Person` or an `Organisation` (subclasses). A CV may refer to the interests of its owner, which might have been explicitly stated by the owner (`hasExplicitInterest`), or might have been inferred by the interaction in the marketplace (`hasImplicitInterest`). When we think of crowd workers, a `CV` may be related to `obtained qualifications`, which are related to competencies. We propose the use of the SKOS vocabulary for competences included in the Europass[7], but any taxonomy about crowdsourcing-oriented skills can be connected in the same way. For each piece of work accomplished, the `CV` connects the relation `hasWorkerExperience` to a new `WorkerExperience`, which consists of information about the way the crowd worker solved the microtasks (e. g. the time the worker invested, whether the requester gave flags or stars in such work, and the engagement of the worker in the complete group of microtasks). When we think of requesters, a CV is related to the `RequesterExperience`, which refers to the work they offer. The CV may have an `Evaluation` associated, which reflects usually a global evaluation connected to a particular `UserAccount` (e. g. the global reputation of a worker in a marketplace). We align our `CV` class to the CV class in the ResumeRDF vocabulary (with `owl:equivalentClass`).

UserAccount represents the account that an `Agent` may have in a marketplace. It is a SIOC class, to which we associate a role defined in SKOS (`Requester` or `Worker`). Besides the geographical information related to a `UserAccount`, what is relevant for the `CV` is the relation between the `UserAccount` and the `Marketplace` where the account belongs to. A crowd worker may have several accounts (one per marketplace), which are described with a `username`, the `language`(s) spoken by the owner and its `creationDate`.

Qualification refers to the achievement that determines whether an agent (usually a crowd worker) has the required knowledge on a particular topic. Qualifications–which are obtained through qualification tests–can be specified as requirements of microtasks (`hasQualification`), to restrict the set of crowd workers who may accomplish the microtasks. Requesters can write their own tests or reuse the questions provided by marketplaces. In the Crowd Work CV, qualifications may be defined with a textual `description`, a `URL` with an deployed example and a `name`. This class may be extended in the future if categories of qualifications are defined (e. g. a subclass could be language qualifications).

MasterMicrotask is a set of `Microtasks` grouped by the same structure, description and configurations. Usually microtasks are generated applying a template (for the UI and other crowdsourcing settings). Templates are combined with data and the `Marketplaces` convert these into specific microtasks. We have collected a set of common microtask purposes (e. g. from the task templates published by CrowdFlower)

[3] Implementation of the Crowd Work CV data model: `https://github.com/criscod/CrowdWorkCV/tree/master/ontology`
[4] FOAF vocabulary `http://xmlns.com/foaf/spec/`
[5] SIOC Core Ontology Specification`http://www.w3.org/Submission/sioc-spec/`
[6] ResumeRDF `http://www.w3.org/wiki/ResumeRDFOntology`
[7] DISCO `http://disco-tools.eu/disco2_portal/`

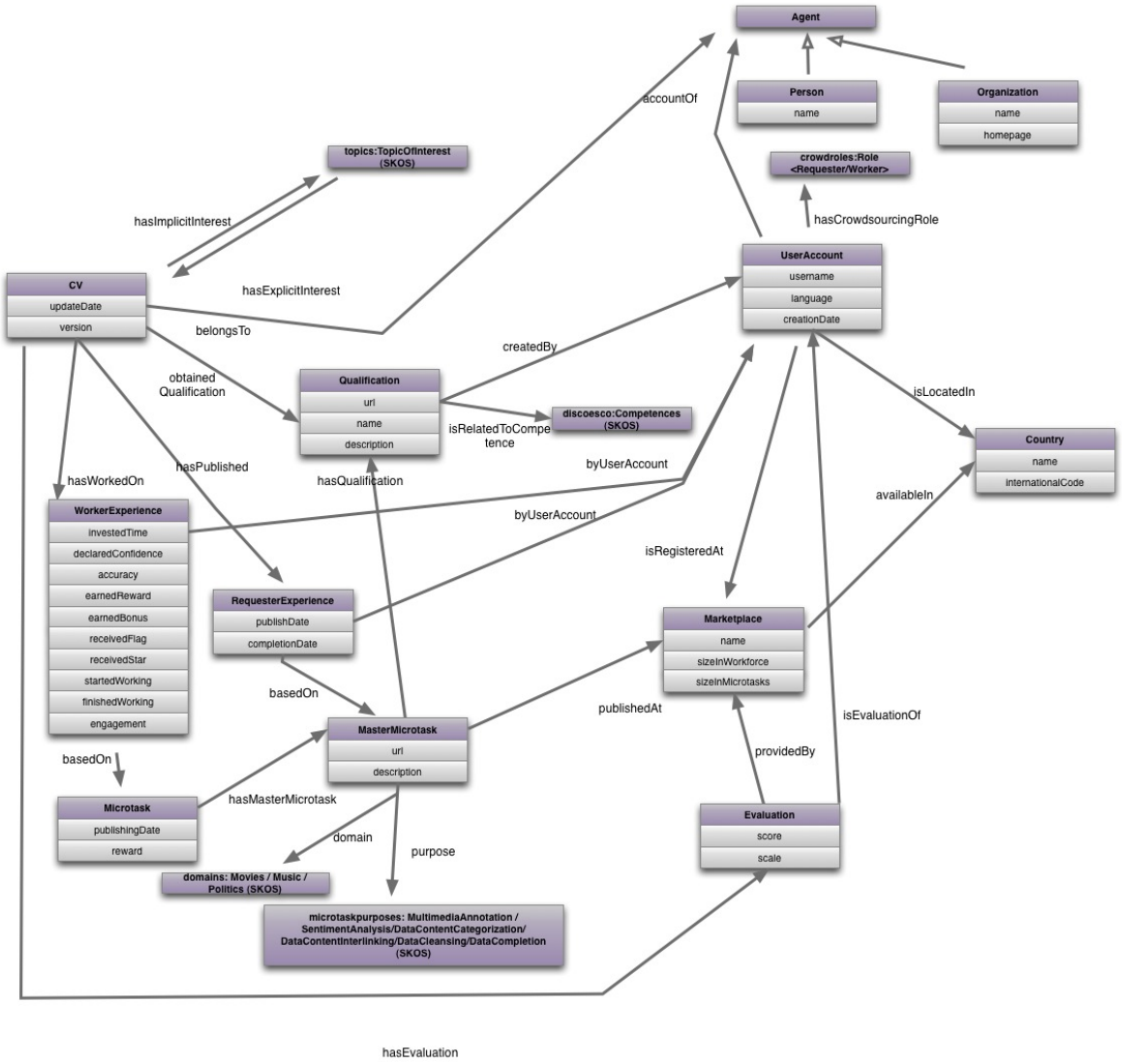

Figure 6.1: Overview of the Crowd Work CV ontology. With Crowd CV it is possible to describe agents, their user accounts, CVs, qualifications, work experiences, microtasks and their master microtasks, marketplaces.

and defined a SKOS vocabulary with these. The taxonomy can be further extended. In the same way, we have included in a SKOS vocabulary some examples of possible domains. New domains and purposes that marketplaces or requesters may define can also be included.

Microtask represents the particular instances of `MasterMicrotasks`. The specific unit of work that crowd workers need to solve. `WorkerExperience` is related to `Microtask`, since the information associated to the WorkerExperience (e. g. accuracy, invested time) is based on (`basedOn`) the results obtained in the microtasks. A `Microtask` is related to the `MasterMicrotask` which from it originated (after combining a template with data).

Marketplace represents the crowdsourcing platform containing a Website where microtasks are offered and accomplished. Such platforms provide support for both requesters (for creating the microtasks, defin-

ing basic restrictions on who to accept in their microtasks, monitoring the evolution of the work, and obtaining the crowdsourced results) and crowd workers (for browsing available microtasks, acquiring qualifications, submit their work, monitoring their activity in the marketplace and sending their feedback). `Marketplaces` may be described by a `name` and their `sizeInMicrotasks` and `sizeInWorkforce` to have some statistical information about them.

Evaluation reflects the assessment of an `Agent` in a `Marketplace`. It is generally described by a `score` within a `scale`, but it could easily be extended, for example with new intermediate properties that following the criteria suggested by Turkopticon express that a `Requester` is evaluated by its communicativity, generosity, fairness and promptness[8]

Figure 6.2 shows an excerpt of the serialisation of the motivational scenario. The complete data can be found at the GitHub repository[9]

```
ex:rex a cwcv:RequesterExperience;
    cwcv:byUserAccount ex:acc1;

ex:cv1 cwcv:hasRequesterExperience
ex:rex;

ex:mm1 cwcv:publishedAt ex:ClickSense;
    cwcv:publishedAt ex:GetPaid;
    cwcv:purpose
microtaskpurposes:SentimentAnalysis

ex:cv2 cwcv:obtainedQualification
          ex:q1;
ex:q1 a cwcv:Qualification;
    cwcv:name "Spanish A1";
    cwcv:isRelatedToCompetence
       disco:Capability1;

ex:mm2 a cwcv:MasterMicrotask;
    cwcv:hasQualification ex:q1;
    cwcv:publishedAt ex:Neobux;

ex:cv2 cwcv:hasWorkerExperience ex:mex1;
ex:mex1 a cwcv:WorkerExperience;
    ex:mex cwcv:basedOn ex:m1;
    cwcv:byUserAccount ex:acc2;
    cwcv:accuracy "0.9";
ex:m1 a cwcv:Microtask;
    cwcv:hasMasterMicrotask ex:mm1;
```

Figure 6.2: Crowd Work CV data to describe the work accomplished in marketplaces. For each work done or published an experience is created.

6.3.2 Ontology verification

In order to ensure that we are following best practices in ontology engineering, we validated our ontology with the OOPS! pitfall scanner[10], which considers a list of 40 common pitfalls in ontology specifications. Except for the imported concepts and properties from other ontologies, we ensured that we do not have important nor critical piftalls.

We also verified the fulfillment of the aforementioned Crowd Work CV requirements: the main elements of the Crowd Work CV ontology refer to **domain-independent** objects in crowdsourcing systems (e. g. microtasks, user accounts and marketplaces). The SKOS vocabularies connected to the core of the Crowd Work CV ontology, which express the purpose of microtasks or the domain, are responsible for bringing the specific knowledge domain into the CV data. Along the same lines, the ontology elements are general enough to be used in **different marketplaces**. For instance, the overall evaluation of a worker in a marketplace or the qualifications, do not refer to particular evaluation schemes that MTurk or Clickworker have—which might be different from other marketplaces. The **semantic and syntactic interoperability** of the Crowd Work CV data is achieved with the use of the OWL ontology language. Furthermore, the Crowd Work CV ontology can easily be extended by defining subclasses (e. g. subclasses of qualifications), subproperties, or adding new relations between existing and new

[8] Turkopticon's evaluation criteria `http://turkopticon.ucsd.edu/help`

[9] Example of generated Crowd Work CV: `https://github.com/criscod/CrowdWorkCV/tree/master/ontology`

[10] `http://oeg-lia3.dia.fi.upm.es/oops/index-content.jsp`

ontology concepts. The SKOS vocabularies can also be easily extended in order to have for example, a broader catalogue of microtask purposes. The Crowd Work CV ontology is compliant with existing **standard traditional CV information**, describing the particular instances of work experience, the educational achievements (in our case qualifications) and related skills (more details on the comparison can be found in the GitHub repository[11]).

6.4 Related Work

Several authors have proposed new methods for matching crowd workers and tasks in crowdsourcing environments. Khazankin et al. (2011b) defined a framework for selecting suitable crowd workers to solve a task based on skill requirements attached to tasks, the availability workers report they have, and the skills workers have. Gagan Goel and Singla (2013) introduced a method for assigning tasks to workers, which analyses both skills and costs. Difallah et al. (2013b) implemented in a Facebook App a recommendation strategy that pushes suitable tasks to users based on information extracted from their Facebook profiles and previously accomplished HITs, following various assignment strategies (i. e. category-based, text-based, and graph-based). These approaches do not offer a shareable and reusable description of worker expertise that could be used across-platforms. ul Hassan et al. (2013) proposed the SLUA ontology for matching users and actions in crowdsourcing scenarios. While the authors raised the problem of lacking interoperability between platforms aligned to our initial proposition, their approach has a different focus: they describe tasks, users, rewards and capabilities primarily for routing. In contrast, our goal is to gather more information and be able to share it as a means to recognition for work. We also consider microtasks, marketplaces, qualifications and requesters' information. Moreover, our data will lead to a workflow for building CV summaries out of large sets of RDF triples. ResumeRDF[12] is an RDFS vocabulary to express information of Curriculum Vitae, including personal details, attended courses, skills and work experience. Celino (2012) proposed the Human Computation ontology, which enables the annotation of crowdsourced data and is mapped to the Provenance Ontology. These data models share some common concepts with ours but do not cover all the crowdsourcing-specific domain required in a Crowd Work CV.

6.5 Conclusions and Future Work

Because microtask crowdsourcing is a social evolving ecosystem with humans on both sides, who invest time and money with a purpose, we need to define methods that satisfy the needs and expectations of all involved agents. We have presented Crowd Work CV, an approach for modelling and sharing knowledge about crowd work experience across different marketplaces, which could facilitate a fruitful requester-crowd worker interaction in microtask marketplaces and weave relations of trust. Our approach would considerably enrich the way reputation and credentials are managed in the current crowd workplace. The Crowd Work CV would also encourage job specialisation policies in microtask crowdsourcing.

Future work will focus on the development of the infrastructure of the Crowd Work CV data management system. An interesting area we would like to investigate is the automatic generation of Crowd Work CV summaries out of large sets of Crowd Work CV RDF triples.

6.6 Acknowledgements

The research leading to these results has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 611242 SENSE4US

[11] `https://github.com/criscod/CrowdWorkCV/blob/master/ontology/EuropassLinkedIncomparison.txt`

[12] ResumeRDF `http://rdfs.org/resume-rdf/`

Chapter 7

Conclusions

Data integration is one of the data management tasks that received the most attention from academia and industry over time. Various technologies were invented to address different specific needs, that emerge in either closed, medium-size, high-quality data scenarios, or in open, large-scale Web scenarios with incomplete/inaccurate data. Semantic Web technologies opened new horizons for data integration problems that entail many, dynamic and truly heterogeneous sources in the context of the Web. These technologies provide a very flexible mechanism to represent links (at a schema- and data-level), and the detailed machine-readable descriptions of entities often facilitate the discovery of links. Nevertheless, as acknowledged by many, the field in continuous progress, and there is still much to investigate and innovate.

Despite the significant improvement of computational methods for link discovery, including techniques like deep learning, humans still play a highly valuable role in ensuring that this task is solved effectively. The creation of ground truth datasets require human agreement and labeling; decisions about the datasets to be linked, the ontologies to be mapped, as well as the type of relationships to be defined, need to be initiated and iteratively revised by humans; complex alignments and low-quality data (e. g. incomplete or inaccurate) often require additional actions that hardly follow a pattern. Hence, it is of utmost importance to facilitate workflows in which humans and machines are combined, so as to ensure a resulting high-quality linked data.

Throughout the body of work produced in the context of this thesis, we conducted empirical research to study how to design solutions for several sub-problems relevant to the process of link quality management, where we (i) *involve* humans to extend machine computation, (ii) *support* humans in the process of assessing and improving links, and (iii) *analyze* humans' behavior to improve the community management in a linked data peer-production system.

The contributions of this thesis can be summarized as follows:

- We provide an implementation of a prototype that given ontologies as input automatically creates microtasks and publishes them into a crowdsourcing marketplace, retrieves responses from the crowd, and produces an aggregated alignment solution manually curated by crowd workers as a result.
- We conducted a feasibility study on crowdsourcing ontology alignment with various ontologies and microtask designs.
- We defined and implemented a set of novel diversity-based measures to facilitate data publishers assess the information gain enabled by existing links from their datasets to other datasets.
- We provided an evaluation of the validity of these measures, by applying them over 35 real-world linked datasets.
- We ran a data-driven longitudinal analysis of Wikidata editors, to identify differences between standard and power editors in terms of the evolution of their editing behavior over time.

- We developed and evaluated predictive models to predict whether editors will become standard/power editors in terms of volume of edits and lifespan.
- We designed and implemented an ontology to capture the history of work across marketplaces, and enable a seamless exchange of this information between crowd workers and requesters.

Taken together, the results of this research suggest that while the link discovery problem has been scientifically studied in-depth, and for a long period of time, there is still room for improving human-machine workflows for link quality assessment.

With the existence of crowdsourcing platforms like Amazon Mechanical Turk, peer-produced Linked Data projects like Wikidata, and other participatory technologies, we have an unprecedented opportunity to augment computational methods with large-scale human computation, regardless of its nature —expert, laypeople, or a combination thereof. However, in order to leverage the potential of these technologies, it is important that we try to address current participation inequalities by trying to involve and coordinate efficiently as many people as possible, taking into account that people's diversity (in e. g. knowledge, experience, preferences, cultural background and personality) may noticeably enhance the quality of the result.

Moreover, if we expect to truly widen the adoption of these technologies and at the same time maintain certain data and process quality standards, we need to build appropriate interfaces for different parts of the population. These interfaces have the two-fold goal of (i) ensuring a common understanding among all people involved in data modeling and data query constructs, and (ii) tightly integrating human-human and human-machine actions. For example, Wikidata currently has the challenge of dealing with `rdfs:subClassOf (wdt:P279)` and `rdf:type (wdt:P31)` statements that were incorrectly created by editors who did not notice the semantic difference between these two predicates.

Semantic Web technologies open countless opportunities to link resources on a dynamic, distributed, and heterogeneous Web. Traditionally, data has been semantically linked by identity, often by publishers with no predefined specific *consumption purpose*, who would like to make their datasets more easily discoverable. While the result of this practice (e. g. the materialization of millions of `owl:sameAs` links (Beek et al., 2018)) is already very useful, Semantic Web links can be an asset for further consumption use cases, including entity-based search and querying. Given that datasets are becoming increasingly discoverable with the appearance of Web portals and dataset search engines that operate as indexes of the global dataspace, it is important that we define further linking optimization criteria.

7.1 Limitations

Despite having carefully designed research methodologies that adhere to scientific standards throughout our work, there are (as it is common for any piece of research) several limitations that can be identified:

First, in our investigation related to the application of microtask crowdsourcing to the problem of ontology alignment we did not evaluate our approach on complex alignments as defined by Thiéblin et al. (2019). While many of our solution constructs would be directly applicable to complex alignments (e. g. the general task design, the quality assurance mechanisms that we implemented, the human-machine workflow), this type of alignments would presumably increase the difficulty of the human computation task. Hence, we expect that the concrete human-data interaction design would need to be adjusted, so as to collect effective input from the crowd. However, we do not identify a reason to reject its feasibility a priori, before an empirical investigation, because related work showed that crowdsourcing can also be effective in tasks with a higher degree of difficulty, as long as relevant information is conveyed clearly. For example, Mortensen et al. (2013) ran successful experiments on crowd-powered ontology verification in the biomedical domain. It is also important to note that at the time of our research, complex alignments were not included in the benchmarking of ontology alignment algorithms.

Second, as indicated in (Chapter 4) we validated the link assessment measures that we defined with real-world datasets, and inspected the presence of gain heterogeneity among entities of the same dataset. While these findings prove that there exists a niche for such measures that no state-of-the-art work manages to cover, at the moment we lack evidence about the adoption of these measures and their usefulness to make data publishers update their interlinking for the better (e. g. correcting semantically inaccurate links that can be spotted as an

outlier gaining a suspiciously large amount of information; trying to homogenize or increase the information gain that entities have in a dataset). While it would be desirable to extend this research with a new quantitative and qualitative analysis along these lines, we decided to leave it out of the scope of this thesis, due to the difficulty of engaging a considerable number of data publishers in such an experiment.

Third, in our study about Wikidata editing behavior over time (Chapter 5), one of the dimensions considered in the prediction model that we implemented was the diversity of types of tasks accomplished by editors. The classification of tasks that we employed —distinguishing between addition, update, and deletion of item data— did not provide a very meaningful indication of the type of user that editors would become over time (i. e. with long/short lifespan, and with high/low volume of edits). While this activity classification was aligned with the state of the art, alternative classifications may possibly be more distinctive of editing behavior evolution. One could define a classification with finer granularity, differentiating between different elements in the data model (e. g. triple statements, qualifiers, references, etc.), or alternatively, one could include the edits done in other Wikidata namespaces[1], to include user, projects, schema and templating pages, as well as talk pages. Moreover, given that Wikidata tools and bots have been increasingly used in recent years, one could also extend the study to consider (semi-)automated edits. However, while doing so, one would need to prepare the data in order to account for the large difference that it exists in terms of volume between the two sets of edits (i. e. automated and non-automated edits).

7.2 Outlook

I identify several directions that the aforementioned work could lead to:

Extending Link Quality Assurance Processes with Further Link Quality Dimensions Link discovery algorithms optimize for link accuracy. However, as mentioned earlier in this thesis, in order to guarantee high link quality it is essential to expand the iterative link quality assurance process with other target dimensions, even if that entails applying an ensemble of tools that address different data/link quality dimensions. The measures introduced in Chapter 4 provide a means to assess one of these extra dimensions —the extent to which links contribute to information retrieval principles in terms of gaining information at an entity level. Put into practice, these measurements can encourage the creation of new links with not only identity predicates, but also domain-specific or more general predicates. For instance, if we have a very specific dataset with entities that are not typically described in other human knowledge graphs, we could still link to external datasets in order to geo-locate, re-type and document our own entities.

Additionally, there could be further measures that focus on dimensions that are relevant to particular types of use cases. An example could be *data analysis*. Currently, we lack measures that indicate how suitable a collection of (two or more) datasets is for being analyzed together. If we had such measures, we could define new types of links based on them, to indicate the compatibility for data analysis in a varying granularity (i. e. links between datasets and links between sets of entities).

Extending/Adapting Link Discovery and Link Quality Assurance Algorithms with More Human-Friendly and Social Features Link Discovery Frameworks have been traditionally designed such that the initiator (often an expert data publisher) sets at least the source dataset, the target dataset and the type of link predicate, and then sets of candidate pairs are reviewed. However, this paradigm assumes pre-existing knowledge about the datasets. The literature lacks a detailed investigation comparing, for different types of humans and datasets, different ways to (i) allow the user(s) explore the datasets concomitantly, (ii) define link predicates to be used, (iii) review candidate links in sets of different kinds of sequences, and (iv) produce further kinds of bi-directional feedback between the human and the algorithm. Such a study should acknowledge the findings in the literature about how users explore ontological knowledge (Walk et al., 2017) and how they edit in collaborative ontology engineering environments (Walk et al., 2015b), when they have the freedom to

[1] Wikidata namespaces `https://www.wikidata.org/wiki/Help:Namespaces`

decide what and how to edit, like in Wikidata. Guiding users in this collaborative or individual process —we hypothesize— could lead to a higher volume of accurate and enriching links.

Extending Large-Scale Human-Machine Solutions to Data Science Large-scale human computation has been widely investigated in the field of data management (Doan et al., 2011; Marcus and Parameswaran, 2015). With the considerable increase of available data, further human input is also required to design data science processes, as well as interpret and organize their corresponding results. New initiatives around the fields of human-centered and participatory data science call for methods that involve new types of citizens into data science pipelines, support data scientists in their quest for studying data thoroughly, and ensure basic human-data needs such as explainability. We foresee this research field to be blooming in the next years. In fact, some of these ideas have been leveraged into an ongoing research project called CrowdAlytics funded by the Swiss National Science Foundation, and we are currently investigating them.

Further Analysis and Intervention for Retention Management in Peer-Production There is a clear extension of our work in Chapter 5 that invites to *identify* further behavioral patterns and habits that help so-called power users have a very productive wiki lifespan, as well as problems that newcomers face before dropping out, albeit willing to contribute. There are studies that show the influence of e. g. reverts on newcomers in Wikipedia (Halfaker et al., 2011), but further qualitative and quantitative studies could shed some light on the reasons why certain users, despite contributing adequately, despair at editing for a long period of time or in high volume (e. g. if they have difficulties in grasping the implicit social structures, they do not find what to edit etc.).

A second natural step to extend this line of research is to create CSCW-based intervention strategies to try to reduce the number of people who drop-out. Such a method for attrition detection should be used in combination with a classification method that determines if editors behave in good faith or not. Existing machine learning techniques like ORES (Halfaker and Geiger, 2020) can help with such classification. The idea is that the behavioral change that should be encouraged in good faith and likely-to-drop users is different from malicious and not-likely-to-drop users. Using bots for this purpose might help in keeping the attention of individuals or collectives of users.

Further Human-Machine Integration Mechanisms Despite all the work on human-machine hybrid algorithms, there is still a lot of research to pursue to better integrate these two sources of knowledge and intelligence, in both the design phase and the execution phase. For instance, given a dataset and a data management problem, it is still hard to systematically quantify a priori the amount and type of human computation needed to improve a target variable (e. g. precision) in that data management problem. So far, empirical pilots are required to have an indication. It would be desirable to predict the increment in quality that humans can trigger in hybrid algorithms, based on properties of the data, the task, and prior statistical knowledge.

Bibliography

Acosta, M., Zaveri, A., Simperl, E., Kontokostas, D., Auer, S., and Lehmann, J. (2013). Crowdsourcing linked data quality assessment. In *The Semantic Web - ISWC 2013 - 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part II*, pages 260–276.

Albertoni, R., Isaac, A., Debattista, J., Dekkers, M., Guéret, C., Lee, D., Mihindukulasooriya, N., and Zaveri, A. (2016). Linked data. W3C Working Group Note.

Albertoni, R., Martino, M. D., and Podestà, P. (2015). A linkset quality metric measuring multilingual gain in skos thesauri. In Rula, A., Zaveri, A., Knuth, M., and Kontokostas, D., editors, *LDQ@ESWC*, volume 1376 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Albertoni, R. and Pérez, A. G. (2013). Assessing linkset quality for complementing third-party datasets. In *Proceedings of the Joint EDBT/ICDT 2013 Workshops*, EDBT '13.

Alvarez, M. R. (2016). *Computational Social Science: Discovery and Prediction*. Analytical Methods for Social Research. Cambridge: Cambridge University Press.

Ang, L. and Buttle, F. (2006). Customer retention management processes: A quantitative study. *European Journal of Marketing*, 40(1/2):83–99.

Antoniou, G., Groth, P. T., van Harmelen, F., and Hoekstra, R. (2012). *A Semantic Web Primer, 3rd Edition*. MIT Press.

Aranda, C. B., Polleres, A., and Umbrich, J. (2014). Strategies for executing federated queries in SPARQL1.1. In Mika, P., Tudorache, T., Bernstein, A., Welty, C., Knoblock, C. A., Vrandecic, D., Groth, P. T., Noy, N. F., Janowicz, K., and Goble, C. A., editors, *The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part II*, volume 8797 of *Lecture Notes in Computer Science*, pages 390–405. Springer.

Asprino, L., Beek, W., Ciancarini, P., van Harmelen, F., and Presutti, V. (2019). Observing LOD using equivalent set graphs: It is mostly flat and sparsely linked. In *The Semantic Web - ISWC 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings, Part I*, pages 57–74.

Batini, C., Cappiello, C., Francalanci, C., and Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computer Surveys*, 41(3):16:1–16:52.

Batini, C., Scannapieco, M., et al. (2016). Data and information quality. *Cham, Switzerland: Springer International Publishing. Google Scholar*, page 43.

Beek, W., Raad, J., Wielemaker, J., and van Harmelen, F. (2018). sameas.cc: The closure of 500m owl: sameas statements. In Gangemi, A., Navigli, R., Vidal, M., Hitzler, P., Troncy, R., Hollink, L., Tordai, A., and Alam, M., editors, *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*, volume 10843 of *Lecture Notes in Computer Science*, pages 65–80. Springer.

Beek, W., Rietveld, L., Bazoobandi, H. R., Wielemaker, J., and Schlobach, S. (2014). LOD laundromat: A uniform way of publishing other people's dirty data. In Mika, P., Tudorache, T., Bernstein, A., Welty, C., Knoblock, C. A., Vrandecic, D., Groth, P., Noy, N. F., Janowicz, K., and Goble, C. A., editors, *The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I*, volume 8796 of *Lecture Notes in Computer Science*, pages 213–228. Springer.

Beek, W., Schlobach, S., and van Harmelen, F. (2016). A contextualised semantics for owl: sameas. In Sack, H., Blomqvist, E., d'Aquin, M., Ghidini, C., Ponzetto, S. P., and Lange, C., editors, *The Semantic Web. Latest Advances and New Domains - 13th International Conference, ESWC 2016, Heraklion, Crete, Greece, May 29*

- *June 2, 2016, Proceedings*, volume 9678 of *Lecture Notes in Computer Science*, pages 405–419. Springer.
Behkamal, B., Kahani, M., Bagheri, E., and Jeremic, Z. (2014). A metrics-driven approach for quality assessment of linked open data. *Journal of theoretical and applied electronic commerce research*, 9(2):64–79.
Benkler, Y., Shaw, A., and Hill, B. M. (2015). Peer production: A form of collective intelligence. *Handbook of Collective Intelligence*, 175.
Berners-Lee, T. (2006). Linked data - design issues. Accessed 2020-02-2020.
Berners-Lee, T., Hendler, J., Lassila, O., et al. (2001). The semantic web. *Scientific american*, 284(5):28–37.
Bernstein, A., Hendler, J., and Noy, N. (2016). A new look at the Semantic Web. *Communications of the ACM*, 59(9):35–37.
Bernstein, M., Karger, D., Miller, R., and Brandt, J. (2012). Analytic Methods for Optimizing Realtime Crowdsourcing. *CoRR*, abs/1204.2995.
Bernstein, M., Little, G., Miller, R., Hartmann, B., Ackerman, M., Karger, D., Crowell, D., and Panovich, K. (2010). Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on user interface software and technology*, pages 313–322.
Berrueta, D., Brickley, D., Decker, S., Fernández, S., Görn, C., Harth, A., Heath, T., Idehen, K., Kjernsmo, K., Miles, A., Passant, A., Polleres, A., Polo, L., and Sintek, M. (2007). Sioc core ontology specification. W3c member submission, W3C.
Bizer, C., Heath, T., and Berners-Lee, T. (2011). Linked data: The story so far. In *Semantic services, interoperability and web applications: emerging concepts*, pages 205–227. IGI Global.
Bock, J., Rudolph, S., and Mutter, M. (2012). More than the sum of its parts - holistic ontology alignment by population-based optimisation. In Lukasiewicz, T. and Sali, A., editors, *FoIKS*, volume 7153 of *Lecture Notes in Computer Science*, pages 71–90. Springer.
Borst, W. N. (1997). *Construction of Engineering Ontologies for knowledge sharing and reuse*. PhD thesis, University of Twente, Enschede.
Brickley, D., Burgess, M., and Noy, N. (2019). Google dataset search: Building a search engine for datasets in an open web ecosystem. In *The World Wide Web Conference*, pages 1365–1375.
Brickley, D. and Miller, L. (2004). FOAF Vocabulary Specification. Namespace Document 2 Sept 2004, FOAF Project. http://xmlns.com/foaf/0.1/.
Carlson, S. and Seely, A. (2017). Using openrefine's reconciliation to validate local authority headings. *Cataloging & Classification Quarterly*, 55(1):1–11.
Celino, I. (2012). Human computation ontology. `http://swa.cefriel.it/ontologies/hc.html`. Accessed 2020-02-2020.
Christen, P. (2012). *Data Matching - Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Data-Centric Systems and Applications. Springer.
Ciampaglia, G. L. and Taraborelli, D. (2015). Moodbar: Increasing new user retention in wikipedia through lightweight socialization. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, page 734–742, New York, NY, USA. Association for Computing Machinery.
Clow, D. (2013). Moocs and the funnel of participation. In *LAK '13. Third Conference on Learning Analytics and Knowledge*, pages 185–189. New York: ACM.
Community, T. W. (2017). Wikidata item quality. `https://www.wikidata.org/wiki/Wikidata:Item_quality`. Accessed 2020-02-2020.
Cosley, D., Frankowski, D., Terveen, L., and Riedl, J. (2007). Suggestbot: Using intelligent task routing to help people find work in wikipedia. In *IUI'07. Proceedings of the 12th International Conference on Intelligent User Interfaces*, IUI '07, pages 32–41. ACM, New York.
Cox, D. R. (1992). Regression models and life-tables. In *Breakthroughs in statistics*, pages 527–541. Springer.
Cuong, T. T. and Müller-Birn, C. (2016). SocInfo'16. applicability of sequence analysis methods in analyzing peer-production systems: A case study in wikidata. In *Social Informatics*, pages 142–156. Berlin: Springer.
Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., and Potts, C. (2013). No country for old members: User lifecycle and linguistic change in online communities. In *WWW 2013. 22nd International World Wide Web Conference, Rio de Janeiro, Brazil, May 13-17, 2013*, pages 307–318. New York: ACM.
David, J., Euzenat, J., and Scharffe, F. (2011). The Alignment API 4.0. *Semantic Web Journal*, 2(1):3–10.
Debattista, J., Auer, S., and Lange, C. (2016). Luzzu—a methodology and framework for linked data quality assessment. *J. Data and Information Quality*, 8(1).

Debattista, J., Lange, C., Auer, S., and Cortis, D. (2018). Evaluating the quality of the LOD cloud: An empirical investigation. *Semantic Web*, 9(6):859–901.

Demartini, G., Difallah, D. E., and Cudré-Mauroux, P. (2012). ZenCrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *Proceedings of the 21st World Wide Web Conference WWW2012*, pages 469–478.

Difallah, D. E., Catasta, M., Demartini, G., and Cudré-Mauroux, P. (2014). Scaling-up the crowd: Micro-task pricing schemes for worker retention and latency improvement. In Bigham, J. P. and Parkes, D. C., editors, *Proceedings of the Seconf AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2014, November 2-4, 2014, Pittsburgh, Pennsylvania, USA*. AAAI.

Difallah, D. E., Demartini, G., and Cudré-Mauroux, P. (2013a). Pick-a-crowd: tell me what you like, and i'll tell you what to do. In Schwabe, D., Almeida, V. A. F., Glaser, H., Baeza-Yates, R., and Moon, S. B., editors, *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013*, pages 367–374. International World Wide Web Conferences Steering Committee / ACM.

Difallah, D. E., Demartini, G., and Cudré-Mauroux, P. (2013b). Pick-a-crowd: tell me what you like, and I'll tell you what to do. In *Proceedings of the 22nd international conference on World Wide Web (WWW2013)*, pages 367–374.

Difallah, D. E., Demartini, G., and Cudré-Mauroux, P. (2016). Scheduling human intelligence tasks in multi-tenant crowd-powered systems. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, pages 855–865.

Dimou, A., Kontokostas, D., Freudenberg, M., Verborgh, R., Lehmann, J., Mannens, E., Hellmann, S., and de Walle, R. V. (2015). Assessing and refining mappings to RDF to improve dataset quality. In Arenas, M., Corcho, Ó., Simperl, E., Strohmaier, M., d'Aquin, M., Srinivas, K., Groth, P., Dumontier, M., Heflin, J., Thirunarayan, K., and Staab, S., editors, *The Semantic Web - ISWC 2015 - 14th International Semantic Web Conference, Bethlehem, PA, USA, October 11-15, 2015, Proceedings, Part II*, volume 9367 of *Lecture Notes in Computer Science*, pages 133–149. Springer.

Dittus, M., Quattrone, G., and Capra, L. (2016). Analysing volunteer engagement in humanitarian mapping: Building contributor communities at large scale. In *CSCW '16. Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work Social Computing*, pages 108–118. New York: ACM.

Dividino, R. Q., Gottron, T., Scherp, A., and Gröner, G. (2014). From changes to dynamics: Dynamics analysis of linked open data sources. In *Proceedings of the 1st International Workshop on Dataset PROFIling & fEderated Search for Linked Data co-located with the 11th Extended Semantic Web Conference, PROFILES@ESWC 2014, Anissaras, Crete, Greece, May 26, 2014.*

Doan, A., Franklin, M. J., Kossmann, D., and Kraska, T. (2011). Crowdsourcing applications and platforms: A data management perspective. *Proceedings of the VLDB Endowment*, 4(12):1508–1509.

Doan, A., Halevy, A. Y., and Ives, Z. G. (2012). *Principles of Data Integration*. Morgan Kaufmann.

Druck, G., Miklau, G., and Mccallum, A. (2008). Learning to predict the quality of contributions to wikipedia. In *WikiAI'08. Proceedings of the Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy*, pages 7–12. Palo Alto: AAAI Press.

Duhigg, C. (2012). *The Power of Habit: Why We Do What We Do in Life and Business*. Random House.

Elmalech, A., Sarne, D., David, E., and Hajaj, C. (2016). Extending workers' attention span through dummy events. In Ghosh, A. and Lease, M., editors, *Proceedings of the Fourth AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2016, 30 October - 3 November, 2016, Austin, Texas, USA*, pages 42–51. AAAI Press.

Ermilov, I., Lehmann, J., Martin, M., and Auer, S. (2016). Lodstats: The data web census dataset. In *The Semantic Web - ISWC 2016 - 15th International Semantic Web Conference, Kobe, Japan, October 17-21, 2016, Proceedings, Part II*, pages 38–46.

Erxleben, F., Günther, M., Krötzsch, M., Mendez, J., and Vrandečić, D. (2014). Introducing wikidata to the linked data web. In Mika, P., Tudorache, T., Bernstein, A., Welty, C., Knoblock, C., Vrandečić, D., Groth, P., Noy, N., Janowicz, K., and Goble, C., editors, *The Semantic Web – ISWC 2014*, pages 50–65, Cham. Springer International Publishing.

Euzenat, J., Shvaiko, P., et al. (2007). *Ontology matching*. Springer.

Falconer, S. M. and Storey, M.-A. (2007). A cognitive support framework for ontology mapping. In *Proceedings of the 6th International Semantic Web Conference (ISWC)*, pages 114–127.

Färber, M., Bartscherer, F., Menne, C., and Rettinger, A. (2018). Linked data quality of dbpedia, freebase, opencyc, wikidata, and YAGO. *Semantic Web*, 9(1):77–129.

Farda-Sarbas, M., Sarasua, C., and Mueller-Birn, C. (2019a). Interactive poster data quality in wikidata. `https://wikimania.wikimedia.org/wiki/2019:Poster_session/Interactive_Poster_Data_Quality_in_Wikidata`. Accessed 2020-02-2020.

Farda-Sarbas, M., Zhu, H., Nest, M. F., and Müller-Birn, C. (2019b). Approving automation: analyzing requests for permissions of bots in wikidata. In *Proceedings of the 15th International Symposium on Open Collaboration*, pages 1–10.

Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395.

Fokou, G., Jean, S., HadjAli, A., and Baron, M. (2016). RDF query relaxation strategies based on failure causes. In *The Semantic Web. Latest Advances and New Domains - 13th International Conference, ESWC 2016, Heraklion, Crete, Greece, May 29 - June 2, 2016, Proceedings*, pages 439–454.

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

Franklin, M. J., Kossmann, D., Kraska, T., Ramesh, S., and Xin, R. (2011). Crowddb: answering queries with crowdsourcing. In *Proceedings of the 2011 international conference on Management of data*, SIGMOD '11, pages 61–72, New York, NY, USA. ACM.

G. Little, L. Chilton, M. G. and Miller, R. (2009). Turkit: tools for iterative tasks on mechanical turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*.

Gagan Goel, A. K. and Singla, A. (2013). Matching workers expertise with tasks: Incentives in heterogeneous crowdsourcing markets. In *NIPS'13 Workshop on Crowdsourcing: Theory, Algorithms and Applications*.

Gamma, E., Helm, R., Johnson, R., and Vlissides, J. M. (1994). *Design Patterns: Elements of Reusable Object-Oriented Software*. Addison-Wesley Professional, 1 edition.

Gandica, Y., Carvalho, J., and dos Aidos, F. S. (2015). Wikipedia editing dynamics. *Physical Review E*, 91(1):012824.

Gangemi, A. (2005). Ontology design patterns for semantic web content. *The Semantic Web – ISWC 2005*, pages 262–276.

Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., and Schneider, L. (2002). Sweetening ontologies with dolce. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 166–181. Springer.

Gaos, Y., Narayanan, A., Patterson, A., Taylor, J., and Jain, A. (2018). Enterprise-scale knowledge graphs. Accessed 2020-02-2020.

Geiger, R. S. and Halfaker, A. (2017). Operationalizing conflict and cooperation between automated software agents in wikipedia: A replication and expansion of'even good bots fight'. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–33.

Geiger, S. R. and Halfaker, A. (2013). Using edit sessions to measure participation in wikipedia. In *CSCW 2013. Computer Supported Cooperative Work, San Antonio, TX, USA, February 23-27, 2013*, pages 861–870. New York: ACM.

Golshan, B., Halevy, A. Y., Mihaila, G. A., and Tan, W. (2017). Data integration: After the teenage years. In *Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS 2017, Chicago, IL, USA, May 14-19, 2017*, pages 101–106.

Gómez-Pérez, A. (2004). Ontology evaluation. In Staab, S. and Studer, R., editors, *Handbook on Ontologies*, International Handbooks on Information Systems, pages 251–274. Springer.

Gordini, N. and Veglio, V. (2017). Customers churn prediction and marketing retention strategies. an application of support vector machines based on the auc parameter-selection technique in b2b e-commerce industry. *Industrial Marketing Management*, 62:100–107.

Görlitz, O. and Staab, S. (2011). Splendid: Sparql endpoint federation exploiting void descriptions. In *Proceedings of the Second International Conference on Consuming Linked Data-Volume 782*, pages 13–24. CEUR-WS. org.

Greiner, A., Isaac, A., Iglesias, C., Laufer, C., Guéret, C., Lee, D., Schepers, D., Stephan, E. G., Kauz, E., Atemezing, G. A., Beeman, H., Bittencourt, I. I., Almeida, J. P., Dekkers, M., Winstanley, P., Archer, P., Albertoni, R., Purohit, S., and Córdova, Y. (2017). Data on the web best practices. w3c recommendation.

Technical report, W3C. Accessed 2020-02-2020.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, 11(1):1–21.

Gruber, T. R. et al. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition*, 5(2):199–220.

Guarino, N., Oberle, D., and Staab, S. (2009). What is an ontology? In *Handbook on ontologies*, pages 1–17. Springer.

Guéret, C., Groth, P. T., Stadler, C., and Lehmann, J. (2012). Assessing linked data mappings using network measures. In *Proc. ESWC 2012*, pages 87–102.

Guha, R. and Brickley, D. (2014). Rdf schema 1.1. w3c recommendation. Technical report, W3C.

Halevy, A. Y., Franklin, M. J., and Maier, D. (2006). Principles of dataspace systems. In Vansummeren, S., editor, *PODS*, pages 1–9. ACM.

Halfaker, A. and Geiger, R. S. (2020). Ores: Lowering barriers with participatory machine learning in wikipedia. *Proceedings of the ACM on Human Computer Interaction*, 4(CSCW2):148:1–148:37.

Halfaker, A., Keyes, O., and Taraborelli, D. (2013). Making peripheral participation legitimate: Reader engagement experiments in wikipedia. In *CSCW 2013. Computer Supported Cooperative Work, San Antonio, TX, USA, February 23-27, 2013*, pages 849–860. New York: ACM.

Halfaker, A., Kittur, A., and Riedl, J. (2011). Don't bite the newbies: How reverts affect the quantity and quality of wikipedia work. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration, 2011, Mountain View, CA, USA, October 3-5, 2011*, pages 163–172. New York: ACM.

Hall, A., Terveen, L., and Halfaker, A. (2018). Bot detection in wikidata using behavioral and other informal cues. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–18.

Halpin, H., Hayes, P. J., McCusker, J. P., McGuinness, D. L., and Thompson, H. S. (2010). When owl: sameas isn't the same: An analysis of identity in linked data. In *The Semantic Web - ISWC 2010 - 9th International Semantic Web Conference, ISWC 2010, Shanghai, China, November 7-11, 2010, Revised Selected Papers, Part I*, pages 305–320.

Hartig, O. (2013). Squin: a traversal based query execution system for the web of linked data. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, pages 1081–1084.

Hausenblas, M., Troncy, R., Raimond, Y., and Bürger, T. (2009). Interlinking multimedia: How to apply linked data principles to multimedia fragments. In *WWW 2009 Workshop: Linked Data on the Web*.

Heath, T. and Bizer, C. (2011). *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool, 1st edition.

Hernández, D., Hogan, A., and Krötzsch, M. (2015). Reifying rdf: What works well with wikidata? *SSWS@ ISWC*, 1457:32–47.

Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P. F., and Rudolph, S. (2009). OWL 2 Web Ontology Language Primer. W3C Recommendation, W3C.

Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutiérrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., Ngomo, A. N., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J. F., Staab, S., and Zimmermann, A. (2021). Knowledge graphs. *ACM Comput. Surv.*, 54(4):71:1–71:37.

Hogan, A., Umbrich, J., Harth, A., Cyganiak, R., Polleres, A., and Decker, S. (2012). An empirical survey of linked data conformance. *Journal of Web Semantics*, 14:14–44.

Hu, W., Qiu, H., and Dumontier, M. (2015). Link analysis of life science linked data. In *The Semantic Web-ISWC 2015*, pages 446–462. Springer.

Huang, S.-W. and Fu, W.-T. (2013). *Don't Hide in the Crowd! Increasing Social Transparency between Peer Workers Improves Crowdsourcing Outcomes*, page 621–630. Association for Computing Machinery, New York, NY, USA.

Iba, T., Nemoto, K., Peters, B., and Gloor, P. A. (2010). Analyzing the creative editing behavior of wikipedia editors through dynamic social network analysis. *Procedia - Social and Behavioral Sciences*, 2(4):6441 – 6456.

Im, J., Zhang, A. X., Schilling, C. J., and Karger, D. (2018). Deliberation and resolution on wikipedia: A case study of requests for comments. *Proceedings of the ACM on Human-Computer Interaction*, 2.

Ipeirotis, P., Provost, F., and Wang, J. (2010). Quality management on Amazon Mechanical Turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pages 64–67.

Ipeirotis, P. G. and Gabrilovich, E. (2014). Quizz: targeted crowdsourcing with a billion (potential) users. In

Proceedings of the 23rd international conference on World wide web, pages 143–154.

Isele, R. and Bizer, C. (2012). Learning expressive linkage rules using genetic programming. *Proceedings of the VLDB Endowment*, 5(11):1638–1649.

Ismayilov, A., Kontokostas, D., Auer, S., Lehmann, J., Hellmann, S., et al. (2018). Wikidata through the eyes of dbpedia. *Semantic Web*, 9(4):493–503.

ISO (2021). Iso 3166 country codes. `https://www.iso.org/iso-3166-country-codes.html`. Accessed 2020-02-2020.

Juran, J. M. (1989). *Juran on Leadership for Quality: An Executive Handbook*. The Free Press.

Käfer, T., Abdelrahman, A., Umbrich, J., O'Byrne, P., and Hogan, A. (2013). Observing linked data dynamics. In *The Semantic Web: Semantics and Big Data, 10th International Conference, ESWC 2013, Montpellier, France, May 26-30, 2013. Proceedings*, pages 213–227.

Kaffee, L. . A., Endris, K. M., and Simperl, E. (2019). When humans and machines collaborate: cross-lingual label editing in wikidata. In Lundell, B., Gamalielsson, J., Morgan, L., and Robles, G., editors, *Proceedings of the 15th International Symposium on Open Collaboration, OpenSym 2019, Skövde, Sweden, August 20-22, 2019*, pages 16: 1–16: 9. ACM.

Kanke, T. (2019). Knowledge curation work in wikidata wikiproject discussions. *Library Hi Tech*.

Kaptelinin, V. and Nardi, B. (2012). *Activity Theory in HCI: Fundamentals and Reflections*. Morgan & Claypool Publishers.

Kaufmann, N. and Schulze, T. (2011). Worker motivation in crowdsourcing and human computation. In *Proceedings of the AAAI workshop on human computation (HCOMP)*.

Kazai, G., Kamps, J., and Milic-Frayling, N. (2011). Worker types and personality traits in crowdsourcing relevance labels. In Macdonald, C., Ounis, I., and Ruthven, I., editors, *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011*, pages 1941–1944. ACM.

Khazankin, R., Psaier, H., Schall, D., and Dustdar, S. (2011a). Qos-based task scheduling in crowdsourcing environments. In Kappel, G., Maamar, Z., and Nezhad, H. R. M., editors, *Service-Oriented Computing - 9th International Conference, ICSOC 2011, Paphos, Cyprus, December 5-8, 2011 Proceedings*, volume 7084 of *Lecture Notes in Computer Science*, pages 297–311. Springer.

Khazankin, R., Psaier, H., Schall, D., and Dustdar, S. (2011b). QoS-based task scheduling in crowdsourcing environments. In *Proceedings of the 9th international conference on Service-Oriented Computing*, pages 297–311. Springer Berlin Heidelberg.

Kittur, A., Chi, E., and Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. In *Proc. 26th annual SIGCHI conf. on human factors in computing systems*, pages 453–456.

Kittur, A., Nickerson, J. V., Bernstein, M. S., Gerber, E. M., Aaron, S., Zimmerman, J., Lease, M., and Horton, J. J. (2013). The future of crowd work. In *16th ACM Conference on Computer Supported Coooperative Work (CSCW 2013)*, pages 1301–1318.

Knublauch, H. and Kontokostas, D. (2017). Shapes constraint language (shacl). w3c recommendation. Technical report, W3C.

Kontokostas, D., Zaveri, A., Auer, S., and Lehmann, J. (2013). Triplecheckmate: A tool for crowdsourcing the quality assessment of linked data. In Klinov, P. and Mouromtsev, D., editors, *Knowledge Engineering and the Semantic Web - 4th International Conference, KESW 2013, St. Petersburg, Russia, October 7-9, 2013. Proceedings*, volume 394 of *Communications in Computer and Information Science*, pages 265–272. Springer.

Krötzsch, M. and Vrandecic, D. (2011). Semantic mediawiki. In Fensel, D., editor, *Foundations for the Web of Information and Services - A Review of 20 Years of Semantic Web Research*, pages 311–326. Springer.

Kulkarni, A. P., Can, M., and Hartmann, B. (2011). Turkomatic: Automatic recursive task and workflow design for mechanical turk. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '11, page 2053–2058, New York, NY, USA. Association for Computing Machinery.

Laniado, D., Kaltenbrunner, A., Castillo, C., and Morell, M. F. (2012). Emotions and dialogue in a peer-production community: The case of wikipedia. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, WikiSym '12, New York, NY, USA. Association for Computing Machinery.

Law, E. and Ahn, L. v. (2011). Human computation. *Synthesis Lectures on Artificial Intelligence and Machine*

Learning, 5(3):1–121.

Law, E., Yin, M., Goh, J., Chen, K., Terry, M. A., and Gajos, K. Z. (2016). Curiosity killed the cat, but makes crowdwork better. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4098–4110.

Lee, Y. W., Strong, D. M., Kahn, B. K., and Wang, R. Y. (2002). AIMQ: a methodology for information quality assessment. *Information & Management*, 40(2):133–146.

Lenzerini, M. (2002). Data integration: A theoretical perspective. In *Proceedings of the Twenty-First ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '02, page 233–246, New York, NY, USA. Association for Computing Machinery.

Lintott, C. J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., Raddick, J. M., Nichol, R. C., Szalay, A., Andreescu, D., et al. (2008). Galaxy zoo: morphologies derived from visual inspection of galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 389(3):1179–1189.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.

Marcus, A. and Parameswaran, A. G. (2015). Crowdsourced data management: Industry and academic perspectives. *Foundations and Trends in Databases*, 6(1-2):1–161.

Markotschi, T. and Völker, J. (2010). GuessWhat?! - Human Intelligence for Mining Linked Data. In *Proceedings of the Workshop on Knowledge Injection into and Extraction from Linked Data at EKAW*.

Max Schmachtenberg, Christian Bizer, A. J. and Cyganiak, R. (2014). Linking open data cloud diagram. Accessed 2020-02-2020.

McCann, R., Shen, W., and Doan, A. (2008). Matching Schemas in Online Communities: A Web 2.0 Approach. In *18th International Conference on Data Engineering(ICDE)*, pages 110–119.

Michie, S., van Stralen, M. M., and West, R. (2011). The behaviour change wheel: A new method for characterising and designing behaviour change interventions. *Implementation Science*, 6(1):42.

Mortensen, J., Musen, M. A., and Noy, N. F. (2013). Crowdsourcing the verification of relationships in biomedical ontologies. In *AMIA 2013, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 16-20, 2013*. AMIA.

Mottin, D., Marascu, A., Roy, S. B., Das, G., Palpanas, T., and Velegrakis, Y. (2014). IQR: an interactive query relaxation system for the empty-answer problem. In *International Conference on Management of Data, SIGMOD 2014, Snowbird, UT, USA, June 22-27, 2014*, pages 1095–1098.

Müller-Birn, C., Karran, B., Lehmann, J., and Luczak-Rösch, M. (2015). Peer-production system or collaborative ontology engineering effort: What is wikidata? In *OpenSym'15. Proceedings of the 11th International Symposium on Open Collaboration*, pages 20:1–20:10. New York: ACM.

Nentwig, M., Hartung, M., Ngonga Ngomo, A.-C., and Rahm, E. (2017). A survey of current link discovery frameworks. *Semantic Web*, 8(3):419–436.

Nentwig, M., Soru, T., Ngomo, A. N., and Rahm, E. (2014). Linklion: A link repository for the web of data. In *The Semantic Web: ESWC 2014 Satellite Events - ESWC 2014 Satellite Events, Anissaras, Crete, Greece, May 25-29, 2014, Revised Selected Papers*, pages 439–443.

Neto, C. B., Kontokostas, D., Hellmann, S., Müller, K., and Brümmer, M. (2016). Assessing quantity and quality of links between linked data datasets. In *Proceedings of the Workshop on Linked Data on the Web co-located with the 25th International World Wide Web Conference (WWW 2016)*.

Neubert, J. (2017). Wikidata as a linking hub for knowledge organization systems? integrating an authority mapping into wikidata and learning lessons for KOS mappings. In Mayr, P., Tudhope, D., Golub, K., Wartena, C., and Luca, E. W. D., editors, *Proceedings of the 17th European Networked Knowledge Organization Systems Workshop co-located with the 21st International Conference on Theory and Practice of Digital Libraries 2017 (TPDL 2017), Thessaloniki, Greece, September 21st, 2017*, volume 1937 of *CEUR Workshop Proceedings*, pages 14–25. CEUR-WS.org.

Ngonga Ngomo, A.-C. and Auer, S. (2011). Limes - a time-efficient approach for large-scale link discovery on the web of data. In *Proceedings of IJCAI*.

Nicola, A. D. and Missikoff, M. (2016). A lightweight methodology for rapid ontology engineering. *Commun. ACM*, 59(3):79–86.

Nov, O. (2007). What motivates wikipedians? *Communications of the ACM*, 50(11):60–64.

Noy, N., Griffith, N., and Musen, M. (2008). Collecting community-based mappings in an ontology repository.

In *Proceedings of the 7th International Semantic Web Conference*, pages 371–386.

Noy, N. F. and McGuinness, D. L. (2001). Ontology development 101: A guide to creating your first ontology. Technical report, Stanford University.

Noy, N. F. and Musen, M. A. (2003). The PROMPT suite: Interactive tools for ontology merging and mapping. *International Journal of Human-Computer Studies*, 59(6):983–1024.

Oberg, J. (1999). Why the mars probe went off course [accident investigation]. *IEEE Spectrum*, 36(12):34–39.

Obrst, L., Ceusters, W., Mani, I., Ray, S., and Smith, B. (2007). *The Evaluation of Ontologies*, pages 139–158. Springer, New York, NY.

Oleson, D., Sorokin, A., Laughlin., G., Hester, V., Le, J., and Biewald, L. (2011). Programmatic gold: targeted and scalable quality assurance in crowdsourcing. In *Proceedings of the 11th AAAI Conference on Human Computation*, AAAIWS'11-11, page 43–48.

Panciera, K., Halfaker, A., and Terveen, L. (2009). Wikipedians are born, not made: A study of power editors on wikipedia. In *Proceedings of the ACM 2009 International Conference on Supporting Group Work*, pages 51–60. New York: ACM.

Pandian, C. R. (2003). *Software metrics: A guide to planning, analysis, and application*. CRC Press.

Paulheim, H. (2014). Identifying wrong links between datasets by multi-dimensional outlier detection. In Lambrix, P., Qi, G., Horridge, M., and Parsia, B., editors, *Proceedings of the Third International Workshop on Debugging Ontologies and Ontology Mappings, WoDOOM 2014, co-located with 11th Extended Semantic Web Conference (ESWC 2014), Anissaras/Hersonissou, Greece, May 26, 2014*, volume 1162 of *CEUR Workshop Proceedings*, pages 27–38. CEUR-WS.org.

Pavel, S. and Euzenat, J. (2013). Ontology matching: State of the art and future challenges. *IEEE Transactions on Knowledge and Data Engineering*, 25(1):158–176.

Pinto, H. S., Tempich, C., and Staab, S. (2009). Ontology engineering and evolution in a distributed world using DILIGENT. In Staab, S. and Studer, R., editors, *Handbook on Ontologies*, International Handbooks on Information Systems, pages 153–176. Springer.

Piscopo, A. and Community, W. (2016). Rfc data quality framework for wikidata. `https://www.wikidata.org/wiki/Wikidata:Requests_for_comment/Data_quality_framework_for_Wikidata`. Accessed 2020-02-2020.

Piscopo, A., Phethean, C., and Simperl, E. (2016). Wikidatians are born: paths to full participation in a collaborative structured knowledge base. In *HICSS 2017. 50th Hawaii International Conference on System Sciences, Hilton Waikoloa Village, Hawaii, USA, January 4-7, 2017*, pages 4354–4363. AIS Electronic Library (AISeL).

Piscopo, A., Phethean, C., and Simperl, E. (2017). What makes a good collaborative knowledge graph: Group composition and quality in wikidata. In Ciampaglia, G. L., Mashhadi, A. J., and Yasseri, T., editors, *Social Informatics - 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part I*, volume 10539 of *Lecture Notes in Computer Science*, pages 305–322. Springer.

Piscopo, A. and Simperl, E. (2019). What we talk about when we talk about wikidata quality: a literature survey. In Lundell, B., Gamalielsson, J., Morgan, L., and Robles, G., editors, *Proceedings of the 15th International Symposium on Open Collaboration, OpenSym 2019, Skövde, Sweden, August 20-22, 2019*, pages 17:1–17:11. ACM.

Ponciano, L. and Brasileiro, F. (2014). Finding volunteers' engagement profiles in human computation for citizen science projects. *Human Computation*, 1(2):247–266.

Poveda-Villalón, M., Gómez-Pérez, A., and Suárez-Figueroa, M. C. (2014). Oops! (ontology pitfall scanner!): An on-line tool for ontology evaluation. *Int. J. Semantic Web Inf. Syst.*, 10(2):7–34.

Prud'hommeaux, E., Harris, S., and Seaborne, A. (2013). Sparql 1.1 query language. w3c recommendation. Technical report, W3C.

Prud'hommeaux, E., Buil-Aranda, C., et al. (2013). Sparql 1.1 federated query. w3c recommendation.

Quinn, A. J. and Bederson, B. B. (2011). Human computation: A survey and taxonomy of a growing field. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, page 1403–1412.

Raad, J., Beek, W., van Harmelen, F., Pernelle, N., and Saïs, F. (2018). Detecting erroneous identity links on the web using network metrics. In Vrandecic, D., Bontcheva, K., Suárez-Figueroa, M. C., Presutti, V., Celino, I., Sabou, M., Kaffee, L., and Simperl, E., editors, *The Semantic Web - ISWC 2018 - 17th International Semantic Web Conference, Monterey, CA, USA, October 8-12, 2018, Proceedings, Part I*, volume 11136 of *Lecture Notes in Computer Science*, pages 391–407. Springer.

Raad, J., Pernelle, N., Saïs, F., Beek, W., and van Harmelen, F. (2019). The sameas problem: A survey on identity management in the web of data. *CoRR*, abs/1907.10528.

Rashid, M., Torchiano, M., Rizzo, G., Mihindukulasooriya, N., and Corcho, Ó. (2019). A quality assessment approach for evolving knowledge bases. *Semantic Web*, 10(2):349–383.

Rosenberg, L. J. and Czepiel, J. A. (1984). A marketing approach for customer retention. *Journal of Consumer Marketing*, 1(2):45–51.

Ross, J., Irani, L., Silberman, M. S., Zaldivar, A., and Tomlinson, B. (2010). Who are the crowdworkers?: shifting demographics in mechanical turk. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems*.

Rubinstein, J. S., Meyer, D. E., and Evans, J. E. (2001). Executive control of cognitive processes in task switching. *Journal of experimental psychology: human perception and performance*, 27(4):763.

Rula, A. and Zaveri, A. (2014). Methodology for assessment of linked data quality. In *Proceedings of the 1st Workshop on Linked Data Quality co-located with 10th International Conference on Semantic Systems, LDQ@SEMANTiCS 2014, Leipzig, Germany, September 2nd, 2014.*

Ryan, R. M. and Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1):68.

Sarabadani, A., Halfaker, A., and Taraborelli, D. (2017). Building automated vandalism detection tools for wikidata. In Barrett, R., Cummings, R., Agichtein, E., and Gabrilovich, E., editors, *WWW (Companion Volume)*, pages 1647–1654. ACM.

Satzger, B., Psaier, H., Schall, D., and Dustdar, S. (2013). Auction-based crowdsourcing supporting skill management. *Inf. Syst.*, 38(4):547–560.

Schaekermann, M., Goh, J., Larson, K., and Law, E. (2018). Resolvable vs. irresolvable disagreement: A study on worker deliberation in crowd work. *Proceedings of the ACM on Human-Computer Interaction*, 2(154):19.

Schmachtenberg, M., Bizer, C., and Paulheim, H. (2014). Adoption of the linked data best practices in different topical domains. In *ISWC 2014, The Semantic Web - 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I*, pages 245–260. Berlin: Springer.

Settles, B. (2012a). Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114.

Settles, B. (2012b). *Active Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers.

Shadbolt, N., Berners-Lee, T., and Hall, W. (2006). The semantic web revisited. *IEEE intelligent systems*, 21(3):96–101.

Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55.

Shi, F., Li, J., Tang, J., Xie, G. T., and Li, H. (2009). Actively learning ontology matching via user interaction. In *Proceedings of the 8th International Semantic Web Conference ISWC 2009*, pages 585–600.

Shvaiko, P. and Euzenat, J. (2013). Ontology matching: State of the art and future challenges. *IEEE Trans. Knowl. Data Eng.*, 25(1):158–176.

Singer, P., Helic, D., Hotho, A., and Strohmaier, M. (2015). Hyptrails: A bayesian approach for comparing hypotheses about human trails on the web. In *WWW 2015. Proceedings of the 24th International Conference on World Wide Web, Florence, Italy, May 18-22, 2015*, pages 1003–1013. New York: ACM.

Stewart, O., Lubensky, D., and Huerta, J. M. (2010). Crowdsourcing participation inequality: A scout model for the enterprise domain. In *HCOMP'10. Proceedings of the ACM SIGKDD Workshop on Human Computation*, pages 30–33. New York: ACM.

Strohmaier, M. and Wagner, C. (2014). Computational social science for the world wide web. *IEEE Intelligent Systems*, 29(5):84–88.

Suárez-Figueroa, M. C., Gómez-Pérez, A., and Fernández-López, M. (2012). The neon methodology for ontology engineering. In Suárez-Figueroa, M. C., Gómez-Pérez, A., Motta, E., and Gangemi, A., editors, *Ontology Engineering in a Networked World*, pages 9–34. Springer.

Thaler, S., Siorpaes, K., Mear, D., Simperl, E., and Goodman, C. (2011a). Seafish: A game for collaborative and visual image annotation and interlinking. In *Proceedings of the European Semantic Web Conference (ESWC 2011)*, pages 466–470.

Thaler, S., Siorpaes, K., and Simperl, E. (2011b). SpotTheLink: A Game for Ontology Alignment. In *Proceed-*

ings of the 6th Conference for Professional Knowledge Management.

Thiéblin, E., Haemmerlé, O., Hernandez, N., and Trojahn, C. (2019). Survey on complex ontology matching. *Semantic Web*, 11(Preprint):1–39.

Thomas, G., Thompson, G. R., Chung, C., Barkmeyer, E., Carter, F., Templeton, M., Fox, S., and Hartman, B. (1990). Heterogeneous distributed database systems for production use. *ACM Comput. Surv.*, 22(3):237–266.

Toupikov, N., Umbrich, J., Delbru, R., Hausenblas, M., and Tummarello, G. (2009). Ding! dataset ranking using formal descriptions. In *LDOW*.

Tsvetkova, M., García-Gavilanes, R., Floridi, L., and Yasseri, T. (2017). Even good bots fight: The case of wikipedia. *PloS one*, 12(2):e0171774.

Tudorache, T. (2020). Ontology engineering: Current state, challenges, and future directions. *Semantic Web*, 11(1):125–138.

ul Hassan, U., O'Riain, S., and Curry, E. (2013). Slua: Towards semantic linking of users with actions in crowdsourcing. In *CrowdSem*.

Umbrich, J., Hogan, A., Polleres, A., and Decker, S. (2015). Link traversal querying for a diverse web of data. *Semantic Web*, 6(6):585–624.

Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., and Chatzisavvas, K. C. (2015). A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*, 55:1–9.

Vandenbussche, P.-Y., Atemezing, G. A., Poveda-Villalón, M., and Vatant, B. (2017a). Linked open vocabularies (lov): a gateway to reusable semantic vocabularies on the web. *Semantic Web*, 8(3):437–452.

Vandenbussche, P.-Y., Umbrich, J., Matteis, L., Hogan, A., and Buil-Aranda, C. (2017b). Sparqles: Monitoring public sparql endpoints. *Semantic web*, 8(6):1049–1065.

Verhoef, P. C. (2003). Understanding the effect of customer relationship management efforts on customer retention and customer share development. *Journal of Marketing*, 67(4):30–45.

Volz, J., Bizer, C., Gaedke, M., and Kobilarov, G. (2009). Discovering and maintaining links on the web of data. In *International Semantic Web Conference (ISWC)*, Chantilly, VA, USA.

Vrandecic, D. (2010). *Ontology evaluation*. PhD thesis, Karlsruhe Institute of Technology.

Vrandecic, D. and Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Walk, S., Helic, D., Geigl, F., and Strohmaier, M. (2016). Activity dynamics in collaboration networks. *ACM Transactions on the Web (TWEB)*, 10(2):11.

Walk, S., Noboa, L. E., Helic, D., Strohmaier, M., and Musen, M. A. (2017). How users explore ontologies on the web: A study of ncbo's bioportal usage logs. In Barrett, R., Cummings, R., Agichtein, E., and Gabrilovich, E., editors, *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, pages 775–784. ACM.

Walk, S., Singer, P., Noboa, L. E., Tudorache, T., Musen, M. A., and Strohmaier, M. (2015a). Understanding how users edit ontologies: Comparing hypotheses about four real-world projects. In *ISWC 2015. Proceedings of the 14th International Conference on The Semantic Web - ISWC 2015 - Volume 9366*, pages 551–568. Springer-Verlag New York, Inc.

Walk, S., Singer, P., Strohmaier, M., Helic, D., Noy, N. F., and Musen, M. A. (2015b). How to apply markov chains for modeling sequential edit patterns in collaborative ontology-engineering projects. *Int. J. Hum. Comput. Stud.*, 84:51–66.

Wang, R. Y. (1998). A product perspective on total data quality management. *Commun. ACM*, 41(2):58–65.

Wang, R. Y. and Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *J. of Management Information Systems*, 12(4):5–33.

West, R., Weber, I., and Castillo, C. (2012). A data-driven sketch of wikipedia editors. In *WWW 2012. Proceedings of the 21st World Wide Web Conference, Lyon, France, April 16-20, 2012 (Companion Volume)*, pages 631–632. New York: ACM.

Whetzel, P. L., Noy, N. F., Shah, N. H., Alexander, P. R., Nyulas, C. I., Tudorache, T., and Musen, M. A. (2011). BioPortal: Enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. *Nucleic Acids Research (NAR)*, 39(Web Server issue):W541–5.

Wiederhold, G. (1992). Mediators in the architecture of future information systems. *Computer*, 25(3):38–49.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten,

J.-W., da Silva Santos, L. B., Bourne, P. E., et al. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3.

Wulczyn, E., West, R., Zia, L., and Leskovec, J. (2016). Growing wikipedia across languages via recommendation. In *WWW 2016. Proceedings of the 25th International Conference on World Wide Web, Montreal, Canada, April 11 - 15, 2016*, pages 975–985. New York: ACM.

Yapinus, G., Sarabadani, A., and Halfaker, A. (2017). Wikidata item quality labels (data set). `https://figshare.com/articles/Wikidata_item_quality_labels/5035796`. Accessed 2020-02-2020.

Yasseri, T., Sumi, R., and Kertész, J. (2012). Circadian patterns of wikipedia editorial activity: A demographic analysis. *PLoS ONE*, 7(1):1–8.

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., and Auer, S. (2016). Quality assessment for linked data: A survey. *Semantic Web*, 7(1):63–93.

Zhdanova, A. and Shvaiko, P. (2006). Community-driven ontology matching. Technical Report DIT-06-028, Ingegneria e Scienza dell'Informazione, University of Trento.

Ziegler, P. and Dittrich, K. R. (2004). Three decades of data integration - all problems solved? In Jacquart, R., editor, *Building the Information Society, IFIP 18th World Computer Congress, Topical Sessions, 22-27 August 2004, Toulouse, France*, volume 156 of *IFIP*, pages 3–12. Kluwer/Springer.