Towards a Systematic Approach to Sync Factual Data across Wikipedia, Wikidata and External Data Sources

Sebastian Hellmann[1][https://global.dbpedia.org/id/3eGWH], Johannes Frey[1][0000−0003−3127−0815], Marvin Hofer[1][0000−0003−4667−5743], Milan Dojchinovski[1,2][0000−0003−4844−4260], Krzysztof Węcel[3][0000−0001−5641−3160], and Włodzimierz Lewoniewski[3][0000−0002−0163−5492]

[1] Knowledge Integration and Language Technologies (KILT/AKSW)
DBpedia Association/InfAI, Leipzig University, Germany
lastname@informatik.uni-leipzig.de
[2] Web Intelligence Research Group
FIT, Czech Technical University in Prague, Czech Republic
milan.dojchinovski@fit.cvut.cz
[3] Poznań University of Economics and Business, Poland

Abstract. This paper addresses one of the largest and most complex data curation workflows in existence: Wikipedia and Wikidata, with a high number of users and curators adding factual information from external sources via a non-systematic Wiki workflow to Wikipedia's infoboxes and Wikidata items. We present high-level analyses of the current state, the challenges and limitations in this workflow and supplement it with a quantitative and semantic analysis of the resulting data spaces by deploying DBpedia's integration and extraction capabilities. Based on an analysis of millions of references from Wikipedia infoboxes in different languages, we can find the most important sources which can be used to enrich other knowledge bases with information of better quality. An initial tool is presented, the GlobalFactSync browser, as a prototype to discuss further measures to develop a more systematic approach for data curation in the WikiVerse.

1 Introduction

Soon after the creation of Wikipedia in 2001, its editors started to define MediaWiki Templates to visually render factual data next to the Wikipedia article text. Over time the amount as well as the variety in these templates have evolved, especially with regard to Wikipedia language editions (WPLE). There is a schematic heterogeneity between templates with different parameters or properties (e.g. over 30 ways for "birthDate", "dateOfBirth") for the same type of value as well as different syntactical ways and sub-templates to express the value itself ("1879-03-14", "March 14th, 1879"). On top, each Wikipedia language edition defines its own templates which in most cases differ from all other 300 language editions. This heterogeneity vastly impacts accessibility of Wikipedia's factual data as well

as multiplies the curation effort, as the same fact needs to be edited individually and manually for each WPLE.

In order to address the heterogeneity in Wikipedia, two key approaches have been developed. First, the DBpedia Mappings and Ontology, which created an interoperability layer by mapping around 80% of the data in infobox templates for the largest 38 Wikipedia languages to the DBpedia Ontology using a crowd-sourced wiki approach starting in 2009 [5]. Second, the Wikidata project started in 2012. One of its goals was to provide a database for systematically querying and inserting factual data into WPLE templates and therefore only curate each fact once centrally.

In June 2019, the GlobalFactSync (GFS) project, which was jointly funded by the Wikimedia Foundation and the DBpedia Association, started with the goal to sync Wikimedia's factual data across all Wikipedia language editions (WPLE) and Wikidata. Here, *factual data* is defined as a certain piece of information, i.e. data values such as "geo-coordinates", "birthdates", "chemical formulas" or relations such as "starring in movies" or "birthplace" attached to an entity (article in Wikipedia, item in Wikidata) and ideally accompanied with a reference of origin. In particular, such facts are found in Wikipedia's infoboxes with a huge variety in quality across languages as well as in Wikidata's statements.

This paper includes the results of the GlobalFactSync (GFS) project, which set out to tackle this long-standing problem of heterogeneity in Wikipedia's templates in a systematic manner by combining the DBpedia and Wikidata approach and its data as well as by developing tools for effective curation of factual data for Wikipedia and Wikidata.

The paper is structured as follows. Section 2 provides background. We analyzed the problem from a high-level conceptual perspective in Section 3 and created a study about the ontological nature of the data under integration aspects in Section 4. Subsequently, we described the developed prototype GlobalFactSync browser in Section 5 as a promising tool to improve curation. We also analyzed the data in a bottom-up manner in Section 6 to gain a complementary view supplementing the high-level perspective. The semantic analysis in Section 7 concerns exploiting the interoperability of the DBpedia Ontology. We discussed outcomes and concluded in Section 8.

2 Background and Related Work

In the context of data integration there is often a question about comparison of Wikidata and DBpedia. Most literature deals with comparing the data and graphs of both projects which is the common denominator [1]. However, the methodology and goals of each project are very distinct. For "Wikidata - the free knowledge base"[4], the knowledge base is the defining aspect and the single focus. Wikidata selects and uses conventional technologies for their main infrastructure to guarantee stability. DBpedia, on the contrary, is mostly an open

[4] `http://wikidata.org`

innovation platform with the goal to innovate and advance the field of knowledge engineering as a whole, including decentralized Linked Data technologies, information extraction, knowledge-graph-based AI & NLP, data curation & integration, graph databases and ontologies to produce well-engineered open knowledge graphs, scalable solutions and open standards to be used and adopted by the technological communities. In sum, while Wikidata focuses on the knowledge base itself, DBpedia focuses on novel ways to maintain, curate and fuse knowledge graphs with the data being side-product and instrument. The main mission of DBpedia is to establish FAIR Linked Data (FAIR, Findability, Accessibility, Interoperability, Reusability) and provide "Global and Unified Access to Knowledge Graphs", thus unlocking the exponential synergies in networked data. Wikipedia and Wikidata are two very important information nodes in this network. Recent DBpedia milestones to bootstrap FAIR Linked Data were the SHACL standard (to validate data in test-driven manner), databus.dbpedia.org (to manage decentral files), MARVIN (to increase agility and efficiency of extraction workflows), archivo.dbpedia.org (to archive and evaluate the web of ontologies) and global.dbpedia.org (to establish a scalable ID system to discover, access and integrate linked data).

In this paper, we focus on the redundant and complementary data across WPLEs itself also considering Wikidata and external sources. Since 2007, DBpedia maintains the DBpedia Information Extraction Framework (DIEF) that originally extracts facts from Wikipiedia's Infoboxes. In the course of time, DIEF was internationalized [5] and extended with support for over 140 languages; the DBpedia Ontology was created as an integration layer over all the Infobox template parameters; Wikidata and Wikimedia Commons were added to the extraction as well as the text of the articles. While there is always a residual of extraction errors, DIEF can be considered the state-of-the-art tool to squeeze the most and the best information from a huge variety of Wikimedia projects (and also other MediaWiki deployments[3]). Recently, DBpedia automated its extraction using the MARVIN bot that produces 22 billion triples per release using an agile workflow [4] and a variety of novel tests to ensure quality.

Due to space reasons, we refer the reader to [1, 5, 7] for more references on the individual parts.

3 Data Curation Mechanics – Lessons Learned

The Law of Diminishing Returns. The curation of data quality follows the well-known law of diminishing returns [8]. One of its consequences is the Pareto principle, which states that 80% of outcomes is provided by 20% of inputs. The above implies that increasing workforce for data curation yields less and fewer returns for each unit added to a point where overall productivity does not increase anymore. The law is very relevant for establishing and coordinating data projects as it implies that any additional human laborer (be it a volunteer or a contracted laborer) will contribute less than the previous one. Moreover, there are two main influencing conditions in data curation: size and heterogeneity. The

size of the data entails proportional growth of the number of errors, which in turn becomes harder to spot with increasing size of data. DBpedia developed RDFUnit[5] and SHACL[6] in the hope of getting an effective handle on data quality issues in huge datasets. As our experience shows, test-driven engineering is limited by heterogeneity, i.e., it is only efficient to test for systematic errors. Making the tests more fine-granular makes them subject to diminishing returns again. These mechanics are particularly relevant in these setups: (1) The DBpedia mappings of templates to the Ontology showed that 20% of the mappings are responsible for 80% coverage of the data; editing by volunteers dropped after the first 20% with "it became too hard" and "was not worth it" given as the main reasons. (2) Only about 5% of the Freebase data could be integrated easily into Wikidata[9]. (3) The law also applies to Wikipedia's infoboxes (not the article text as it is content, not data) (4) and to Wikidata as a whole and in extension, the scripts and bots feeding Wikidata, which also produce systematic and non-systematic data errors and require maintenance.

Data copy. Copying and replication of data are very usual in data integration scenarios. While creating local instances of the data via caching can help with scalability in *data serving scenarios*, it is the first step towards an inefficient data management in *data integration scenarios*. Each copy triggers duplication of data curation efforts on the receiving side, such as mapping, cleansing, transformation, and testing. The duplication of effort is added on top of the above-mentioned law and often causes data consumers from industry to hire dedicated staff for curating data copied into the enterprise from external sources.

Contributions of our paper. So far, we stated general data curation mechanisms and showed the relevance to the data spaces. In the remainder of the paper, we describe the *sync* approach (via linked data) as an alternative, which might in the future be able to improve the mechanics in the following way: systematic updates with references, no duplication of effort if improvements can be pushed upstream, and a higher degree of innovation and automation. We are also investigating the role of Wikidata to evaluate if it improved any of the mechanics in particular for editing Wikipedia's data.

4 Integratability of the Data: The GFS Challenges

In June 2019, we initially selected several sync targets, i.e. concrete domains to sync. We quickly noticed that integration and syncing poses various difficulties for each domain and identified the following four criteria, which we used to classify the domains (see Table 1). The full description can be found here[7].

1. *Ambiguity* – Are the two entities identical, i.e. the same? For each entity, the degree of how easy it is individuated is assessed. For example, there could be the entity Hamburg (i.e. city in Germany) but there are also places with the

[5] `http://rdfunit.aksw.org`
[6] `https://www.w3.org/TR/shacl/`
[7] `https://meta.wikimedia.org/wiki/Grants:Project/DBpedia/GlobalFactSyncRE/SyncTargets`

Table 1. Initial study of selected sync targets difficulties (- - - - - most challenging)

Sync Target	Ambiguity	Property Variability	Reference	Normalization
NBA Players	-	- - - - -	- - -	- - -
Video Games	- - -	- - -	- - - - -	no issues
Movies	- - -	- - - -	- - -	- - -
Music Albums	- - - -	- - -	-	no issues
Music Singles	- - - - -	- - -	-	no issues
Cloud Types	no issues	no issues	- - - -	- - -
Cars	- - -	no issues	- -	- - -
Companies	- - - -	- - -	- - - -	- - -
Cities	- -	- - - -	-	- - -
Employers	- - - -	- - -	- -	- - -
Geo Coordinates	no issues	- - -	no issues	- -

name Hamburg located in the United States. While Wikidata Sitelinks try to tackle this issue across WPLEs, this can be still very challenging when integrating external datasets (e.g. company datasets). Even with Wikidata, (granularity) mismatches in the Wikiverse occur. For example, there is a single German article for the "High Voltage Album" containing two Infoboxes for two versions from 1975 and 1976 while it is linked to the version of 1976 in English and French (both having 2 separate articles for the two albums).

2. *Fixed vs. varying property* describes the rigidness of the methodology used to capture the value, which may change w.r.t. time, nationality or other factors. Normally, the value of birth date is fixed and standardized. A census may record a population count using a different counting method or area each year, resulting in a variety of "correct" but different values.
3. *References* – Depending on the entity's identity check and the property's fixed or varying state, the reference might vary. The criteria judge the accuracy of available references, i.e. whether they could be authoritative or accurate, and also whether they are machine readable. The study is not complete here and there may be more undiscovered sources.
4. *Normalization and conversion of values* – Depending on the language of the article, some properties have varying units (currency, metric vs. imperial system, etc.) or time zones. In order to correctly check the sync state and compare the values these have to be normalized first.

5 Current Sync Approach

5.1 Data Flow

The DBpedia RDF releases are based on monthly Wikimedia XML dumps for Wikidata and over 140 WPLEs. With the help of DBpedia mappings from Wikidata properties as well as from Wikimedia template parameters to DBpedia Ontology properties, both properties and values from the Wikipedia dumps can be extracted by DIEF in a mostly normalized fashion (see Fig. 1). The outcome of

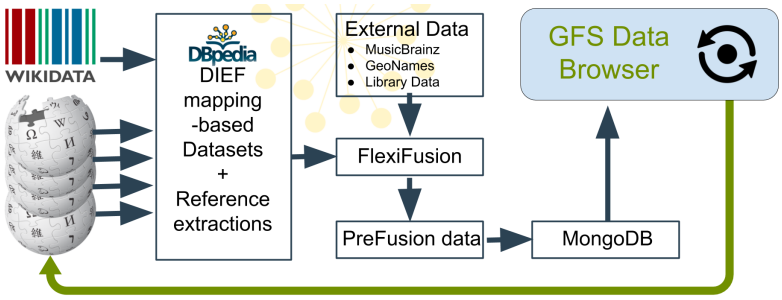

Fig. 1. GFS dataflow: dataflow in black, curation flow in green

these extractions is released in the so-called mapping-based datasets. Using the `InfoboxReferencesExtractor`[8] references can be extracted from Wikitext. As part of the GFS project the DIEF Infobox extractor was extended with the logic of the reference extraction algorithm to enrich the mapping-based facts with Wikipedia references. Additionally, external sources like e.g. German National Library (DNB) and Dutch Library (KB), MusicBrainz and Geonames were partially integrated as a first step to include external linked data sources.

These data sources are then loaded into the DBpedia FlexiFusion [2] pipeline. In a nutshell, FlexiFusion normalizes the identifiers of all entities by using DBpedia Global IDs. Every id represents a cluster of entities standing for the same thing. The clusters were derived by computing connected components based on RDF link sets (`owl:sameAs`) and Wikidata sitelinks. Additional property mappings in the DBpedia Mappings Wiki[9] can be leveraged to compute clusters and Global IDs also for properties. After normalizing the input datasets using Global IDs, the data from various sources is aggregated (pre-fused) for every *sp*-pair. An sp-pair is a JSON data structure which contains for each entity s (using its global ID) and its property p all values from the various sources including provenance (i.e. a link to the input source file for the value). A further step in FlexiFusion allows to "reduce" data (i.e. remove sp-pairs). The PreFusion data is loaded into a MongoDB for further analysis and to feed the GFS Browser.

5.2 GFS Browser and User Script

The GFS Data Browser[10] (see Fig. 2) shows an aggregated view of all available values (and their sources) for one attribute given any Wikimedia article or any entity URL from an integrated external source. It allows to quickly verify the sync status of multiple sources or to review the values for a given source. For

[8] `https://github.com/Lewoniewski/extraction-framework/blob/master/core/src/main/scala/org/dbpedia/extraction/mappings/InfoboxReferencesExtractor.scala`

[9] `http://mappings.dbpedia.org`

[10] `https://global.dbpedia.org`

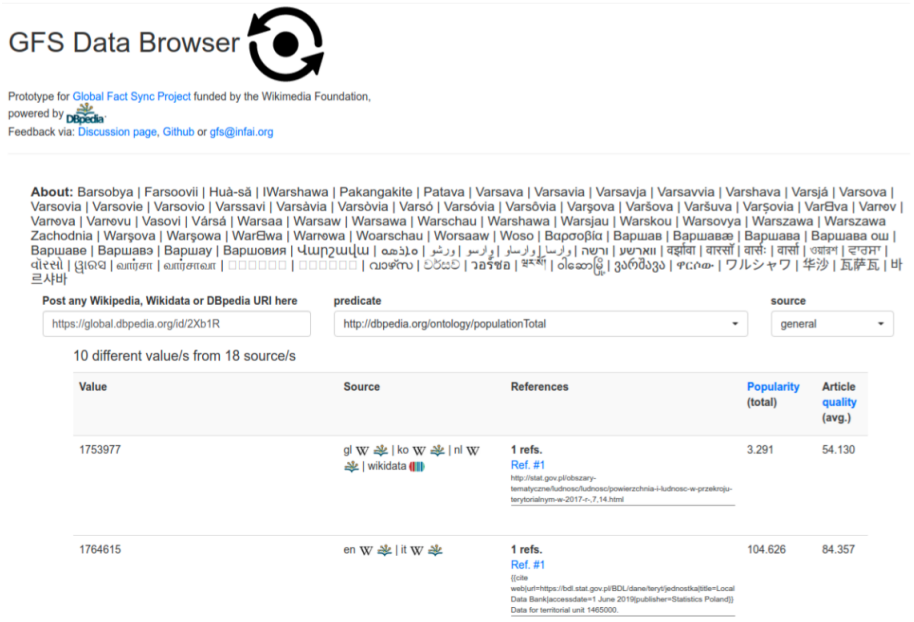

Fig. 2. GFS Browser UI

every fact, provenance in form of links to the original source pages (e.g. Wikidata entities) are displayed. In combination with the InfoboxReferencesExtractor, references for facts extracted from Wikipedia can be displayed. Moreover, quality and popularity measures for the source articles and the attribute in general can be shown. Such measures use a continuous scale, therefore it is possible to compare quality of articles between different WPLEs [6]. A Wikidata or Wikipedia user can quickly review the different values and jump to the Wikidata/Wikipedia pages to edit or fix them.

The GFS User script[11] can be included by any Wikipedia user into their `global.js` file of the MetaWiki. As a consequence, two icons hot-linking to the GFS Data Browser view and the Infobox Reference Extraction microservice will be injected on top of every Wikipedia article and Wikidata entity page. This allows to quickly jump to the GFS browser to view all data available for inclusion in the article.

6 Quantitative Analysis

6.1 Infobox Usage in Wikipedia

Infobox data extracted with DIEF has almost doubled in size since 2016. For the English DBpedia 2016.10 release 52,680,100 raw facts were extracted com-

[11] `https://meta.wikimedia.org/wiki/User:JohannesFre/global.js`

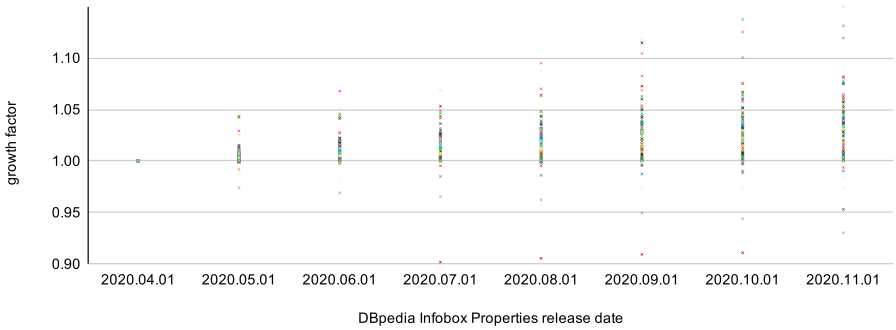

Fig. 3. Infobox extraction growth

pared to 100,101,479 for 2020.11.01. For all languages the total sum raised from 489,805,025 to 725,422,849 facts.

The growth in the number of raw extracted facts can still be observed in the course of this year. Fig. 3 shows that the number of raw facts in the latest 7 monthly releases[12] (between May and November 2020) compared to the reference release in April 2020 is steadily increasing (growth factor greater than 1) for the majority of languages (127 out of 139). The growth factor for the release in November 2020 averaged over all 139 extracted Wikipedia language versions is 1.08 (median 1.03). Major growth was measured for the following WPLEs (abbreviated using Wikipedia.org third-level domain) *arz* (4.29), *vec* (2.99), *ku* (1.55), *nan* (1.41) and *pnb* (1.33), while major decrease was discovered for *la* (0.62), *sco* (0.79), *ast* (0.93), *ga* (0.95) and *sv* (0.97). Also, for the large and popular Wikipedia versions *en* (1.05), *de* (1.06) and *fr* (1.04), we observed an increased number of facts. Supplemental material and raw numbers are available online[13].

We conclude that Infoboxes are still widely used and edited (added, reorganized, completed) and the amount of data in need of sync, as well as the necessary effort to manage it, is growing every day.

6.2 Wikipedia Reference Occurrence Analysis

`InfoboxReferencesExtractor` is a novel extractor for DIEF that was developed in the course of the GFS project. The main goal of the extractor is to get information about references in infoboxes and other templates. Additionally, it uses `CitationExtractor` (another DBpedia extractor) for references that use special citation templates (for example "Cite journal") to extract metadata of the source such as authors, publication date, publisher. The extractor works on a Wikimedia XML dump file with articles, templates, media/file descriptions, and

[12] `https://databus.dbpedia.org/marvin/generic/infobox-properties/`
[13] `https://github.com/dbpedia/gfs/tree/master/report`

Table 2. Number of references extracted from infoboxes/templates

Language	Refs.	AwR	R/AwR	Articles	AR share
en - English	3,254,376	1,042,951	3.120	6,183,371	16.87%
fr - French	1,885,878	271,203	6.954	2,262,520	11.99%
de - German	681,244	164,646	4.138	2,495,604	6.60%
pl - Polish	558,666	256,016	2.182	1,435,232	17.84%
it - Italian	557,423	151,809	3.672	1,645,083	9.23%
es - Spanish	424,345	200,031	2.121	1,637,365	12.22%
ru - Russian	366,420	137,403	2.667	1,672,554	8.22%
sv - Swedish	304,503	101,920	2.988	3,596,319	2.83%
uk - Ukrainian	297,653	177,888	1.673	1,052,352	16.90%
pt - Portuguese	284,412	90,623	3.138	1,045,174	8.67%
nl - Dutch	137,424	62,532	2.198	2,037,601	3.07%

Refs. - number of references in infoboxes and templates; **AwR** - articles with refs. (in infoboxes or templates); **R/AwR** - average number of refs. in AwR.; **Articles** - number of articles in Wikipedia language; **AR share** - share of the articles with refs. in infobox/template

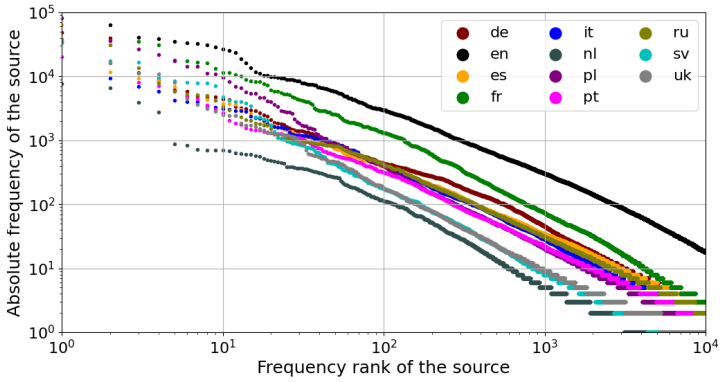

Fig. 4. Frequency distribution of reference URL FQDNs in infoboxes per WPLE

primary meta-pages. The extractor generates two sets of files: *infobox-references* containing references for infoboxes/templates in Wikipedia articles with original parameter names (DBpedia generic extraction), and *mapped-infobox-references* with references of infobox parameters that are mapped to corresponding DBpedia properties (semantic/mapping-based extraction).

The extracted references based on dumps from November 1st, 2020 for 11 WPLEs, which we used for the statistics in the following sections, are available online[14]. The total number of extracted references are summarized in Table 2. Using the generic extraction we found that the most developed WPLE (English) has over 3 million references, which are placed in infoboxes or templates. At the same time, over 1 million articles have at least one parameter using references. Not all sources are cited equally. Some of them are quite popular but there is also a long tail of sources. Figure 4 reveals that the distribution of frequencies of

[14] `http://stats.infoboxes.net/refs/dbpedia/`

Table 3. Top 10 most qualified domain names (FQDN) extracted from references.

FQDN	Counts	Source Type	Topic
ssd.jpl.nasa.gov	757,535	DB source	Astro
web.archive.org	627,886	Archive	*
minorplanetcenter.net	439,590	DB source	Astro
allmusic.com	91,428	Webview[2]	Art
citypopulation.de	87,945	Integrat. Svc.	Gov
webcitation.org	82,178	Archive	*
tvbythenumbers.zap2it.com	79,059	-	Art
census.gov	77,703	Data portal	Gov
spider.seds.org	60,229	Webview[1]	Astro
ned.ipac.caltech.edu	54,352	DB source	Astro

[1]DB source available by other author on request; [2]using comm. data from TiVo

sources in individual languages, visualized on a log-log scale, follows Zipf's law. Further analysis of the frequency along with other measures can help identify the most reliable sources for a specific property of the articles [7]. Although we made use of such quality indicators to support curators in the GFS data browser, these are not in the focus of this paper.

Extraction limitations. In some infoboxes, references are provided not directly near the relevant value but using a separate parameter. For example, in "Infobox settlement" there are special parameters such as "footnotes", "population_footnotes", "population_note". These parameters are often not extracted by DIEF in the mapping-based extraction since they do not have corresponding mappings and they do not have a value to be processed as a property/feature of the subject (e.g. only reference in "<ref>" tag). However, such references are included in generic extraction. An additional challenge is to extract data from references which are described with specific templates without a materialized URL for the source.

6.3 Reference sources

In order to identify primary sources being used for Wikipedia fact spaces, we analyzed the URLs of references from the *infobox-references* files and aggregated them using their fully qualified domain name (FQDN). In Table 3, we report the top 10 cited sources. The sources can be grouped into 4 topics: General Purpose (*), astronomical (Astro), creative artwork (Art), and administrative/governmental domain (Gov). Besides for two web archiving systems and a data portal with diverse files, the remaining sources are either directly rendered from a primary database maintained by the source itself (DB source), present a webview on data derived from a database source, or are integration services from multiple primary data sources (*TVByTheNumbers* was not available at the time of writing). We argue that for these kind of sources a more systematic virtual integration and sync mechanism would have various benefits (correctness, timeliness, efficiency) over the "copy and reference" effort.

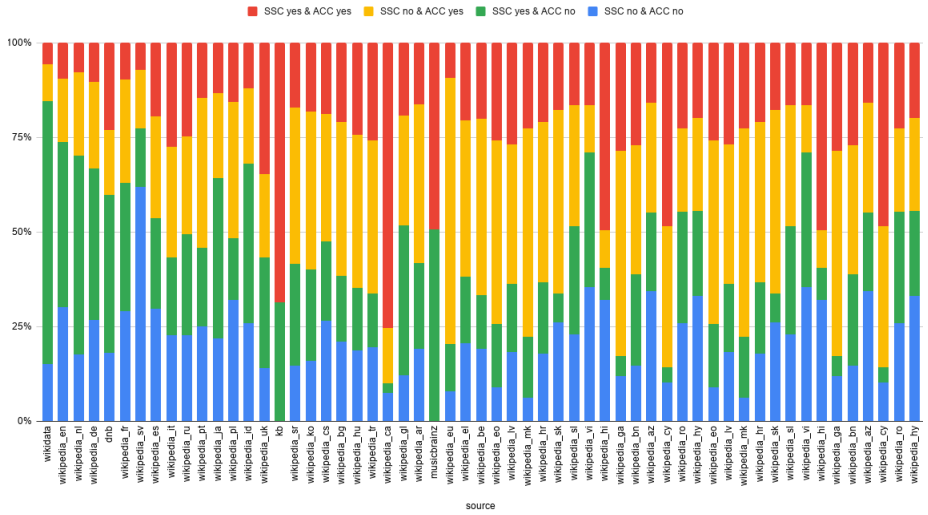

Fig. 5. Distribution of sp-pair sync classification per source: *blue*: value(s) of source are synced / not "challenged"; *green*: property information only in this source (novel or erroneous); *red*: all values from source are not synced, i.e. disjoint; *yellow*: partially synced but also unique value(s) in source, or source is incomplete. The bars are normalized using the total number of sp-pairs per source. The sources are ordered by total number of sp-pairs starting from left with the largest source.

7 Semantic Analysis

7.1 Sources Sync Status

For the sync status analysis of WPLEs, Wikidata and further external sources we used a subset[15] of the PreFusion dataset as of September 2020[16] which only contains properties from the DBpedia ontology having a mapping to a corresponding Wikidata property. Additionally, we reduced the sp-pairs by filtering for pairs where an entity (s) existed in at least one of the WPLEs. This data selection can be considered the factual overlap between all the sources (39 WPLEs having DBpedia Mappings, Wikidata, Musicbrainz, DNB, KB).

In order to study how the different values for one fact (sp-pair) of a source d are in sync w.r.t. the union of all other sources, we defined two binary criteria. The Sole Source Criterion (SSC) is true for an sp-pair of source d if all extracted values contributed from d are only originated in d. The Alternative Choices Available Criterion (ACC) holds if at least one different extracted value from a source other than d is available in the sp-pair. By combining both classifications, we can distinguish 4 different sync states, which are shown in Fig. 5. The sources are unanimous for `no/no`(blue) and agree on these values, interpretable as

[15] `https://databus.dbpedia.org/vehnem/collections/prefusion_p_wikidata/`
[16] `https://databus.dbpedia.org/vehnem/flexifusion/prefusion/2020.09.01`

potentially accurate or consensual information. New unique information is contributed by d in case of `yes/no`(green). Both `no/yes`(yellow) and `yes/yes`(red) have mixed value in need of more elaborate inspection and resolution, whereas `yes/yes`(red) is more polarized and can be interpreted as either complementary beneficial or an erroneous outlier.

With regard to copying data, we can see that on average 21% of equivalent data (blue) is replicated in the Infoboxes across multiple languages. Additional 31% (on average) of the infobox statements have at least a partial redundancy (yellow) but there are also potential contradicting or incomplete values or values out of date (the exact potential cannot be determined due to possible extraction errors and challenges mentioned in Sec. 4 e.g. *variable properties*). The green portion (on average 25%) can be considered novel information that is missing an upstream sync to other sources. This information can be novel and unique with respect to the particular property or attribute (e.g. one source states a death date for a person, while all other sources lack this one) or the entire entity is covered only in this source (e.g. a local company). Finally, on average 23% of sp-pairs have fully disjoint values (red). Besides factual errors or actual complementary data, several factors are likely to contribute to this high number: extraction or normalization problems, incorrect links (from Wikidata or external sources), inaccurate DBpedia mappings or variable properties.

7.2 Reference Analysis with DBO

Using the results of the InfoboxReferencesExtractor (cf. Subsection 6.2), we identified the topics of Wikipedia articles having at least one reference in their infobox. We used the DBpedia Ontology instance types files containing triples generated by the mappings extraction[17] to classify the articles. Table 5 shows the most popular classes with at least one reference. Semantic extraction of references allowed us to identify how often particular infobox parameter in each considered WPLE is supported by a source. To compare this information between languages, parameters of infoboxes were unified to properties using DBpedia mappings. Table 4 shows the top 20 most popular properties with sources for each language version. Despite the popularity, some properties are not supported by references in some languages. Moreover, we took into account the frequency of appearance of references in properties. We found that depending on WPLE we can observe different frequencies of using reference to support the same property.

8 Discussion and Conclusion

Starting the discussion, we would like to mention that we presented many different perspectives on the same huge and complex problem of data curation. Each individual section contributes a certain aspect which we attempt to consolidate *towards* a systematic approach, as good as possible in this discussion section.

[17] `https://databus.dbpedia.org/dbpedia/mappings/instance-types/`

Table 4. Top 20 of the most frequent mapped parameters in each WPLE with at least one reference.

(DBO) Property	total	de	en	es	fr	it	nl	pl	pt	ru	sv	uk
populationTotal	77026	280	7025	5883	59	357	19899	11975	3649	14703	112	13084
nrhpReferenceNumber	59478	-	59478	-	-	-	-	-	-	-	-	-
postalCode	51209	2010	2756	8560	21234	4650	80	5114	305	6462	7	31
foaf:name	43896	1920	28157	2868	792	3119	308	1008	4879	843	-	2
height	36067	827	28101	378	163	2216	294	2972	1115	1	-	-
synonym	35770	-	3606	5314	-	-	-	3424	-	-	23389	37
birthDate	34955	215	31976	3	1120	34	468	1094	1	40	4	-
birthPlace	34318	194	26136	1757	718	85	323	851	838	3197	214	5
Person/height	33745	705	26149	346	155	2203	278	2826	1083	-	-	-
personName	25249	-	6613	-	18581	-	-	-	-	-	-	55
genre	24317	314	11166	2124	3171	13	96	1505	2602	3232	94	-
gross	20924	-	16682	-	-	-	-	-	2423	1792	-	27
budget	20636	-	12766	2577	99	-	563	718	1961	700	313	939
binomialAuthority	20248	-	12777	2662	-	35	-	3804	970	-	-	-
numberOfStudents	18900	2131	14992	384	390	60	162	348	89	231	105	8
numberOfEmployees	18852	7411	6899	787	1356	-	360	561	427	1001	-	50
year	18012	177	56	970	16775	1	16	-	1	8	5	3
revenue	14936	5929	5431	213	850	-	1448	213	32	490	330	-
releaseDate	13831	659	9490	39	1609	1	398	955	10	653	8	9
elevation	12948	195	2063	14	33	7002	9	563	112	2919	37	1

Table 5. Top 20 of the most frequent types in each WPLE with at least one reference for a mapped parameter of the infobox.

DBO Type	total	de	en	es	fr	it	nl	pl	pt	ru	sv	uk
Settlement	209389	33634	63994	-	44499	-	23650	594	6193	-	8457	28368
Album	150848	2997	84274	7893	7867	15258	989	11868	8818	7799	1326	1759
Town	120555	-	32568	-	-	-	-	87948	-	-	39	-
AdministrativeRegion	117142	929	15121	71526	14	-	16	15971	619	7418	3349	2179
PopulatedPlace	89092	-	515	-	-	14953	48	-	4485	25205	2274	41612
SoccerPlayer	79867	2234	37655	351	11340	15508	2564	3515	1738	2846	1134	982
Species	67432	1123	3987	69	-	11036	4562	220	50	6383	40002	-
Film	66916	3313	36346	4075	2451	1358	2096	2260	6953	4834	615	2615
Person	66497	4026	27881	6573	516	20984	267	856	1975	1119	1681	619
Municipality	58028	-	-	4797	-	-	465	-	-	-	8041	44725
City	54282	541	15493	-	25358	-	10	925	5650	-	848	5457
Village	53694	-	48453	-	-	-	-	9	-	5232	-	-
Building	45832	-	41903	991	-	238	111	2290	-	-	153	146
Company	42293	10663	16687	11	4593	1888	781	1252	1733	1918	1732	1035
CelestialBody	41102	-	-	11897	-	943	130	-	-	-	13208	14924
VideoGame	30271	1500	13534	2201	2830	1127	484	1689	1345	3715	1148	698
Politician	27170	-	19365	-	1211	3698	1036	759	929	172	-	-
HistoricPlace	26759	-	26759	-	-	-	-	-	-	-	-	-
River	25618	12136	6578	2	3392	1224	33	966	192	524	23	548
Insect	24833	-	20769	974	-	-	2441	-	649	-	-	-

One of the main lessons learned during the GFS project was that we originally approached the problem from the wrong direction trying to use DBpedia's data to synchronize the WikiVerse, where in fact, data in Wikipedia and Wikidata was collected and copied non-systematically from external sources via many different modalities (manual, scripts, etc.). A major paradigm shift in GFS consisted in starting to exploit the connectivity capabilities of DBpedia (linked data and ontology), which allow to discover and deliver external data from the Linked Open Data (LOD) cloud containing more and more authoritative sources. The LOD cloud has crystallized around DBpedia since 2007 and has grown into the

largest knowledge graph on earth[18]. With the huge size and availability of Linked Open Data and the infrastructure of DBpedia, a unified discovery and automated integration mechanism can be devised, bringing in up-to-date, well-referenced data, with the option to transform Wikidata into a linked data forward proxy and cache (opposite to copy and locally curate, cf. Section 6). In this manner, data in WPLEs could be directly drawn from the source.

The potential of integration is built first of all on the foundation of common identifiers (DBpedia Global IDs). GFS provides the opportunity to at least correlate information in various languages and from various sources. Then, thanks to DBpedia ontology, it is possible to correlate values of parameters, which are first translated to appropriate DBpedia ontology properties. What is not yet translated provides a potential for further mappings.

Comparison of values is supported by assessment of the quality of articles containing it. The last level that we heavily worked on is to bring the external references to the picture. They can further improve the quality of data as questionable values can be verified directly in sources. This also gets in line with requirements of Wikidata (and Wikipedia) to provide references for existing values of data, which very often also evolve over time. Throughout the paper we have provided quantitative evidences.

Using the November 2020 Wikipedia dump, we were able to extract 725.4 million facts from 140 WPLEs as well as 8.8 million references from 11 WPLEs using our novel InfoboxReferenceExtractor, which indicates that there is still a high amount of duplicate effort spent in curating Wikipedia mostly in parallel to Wikidata. We presented an approach (Section 5) that can still greatly improve by integrating the long tail of information from Linked Data.

9 Future Work

Many challenges remain. In Section 6, we showed that the more frequent sources could be effectively integrated, however, the references overall show a long tail distribution, which follow the 20/80 rule (Section 3). Linked Data also provides information and references for this long tail but lack Findability (from FAIR) in its current state, a problem we are currently working on with DBpedia. We have shown that integration and comparison is possible using the (easily extendable) DBpedia Ontology as a pivot point in Section 7 and were able to pinpoint the consensus (blue) and new information (green) between sources. The ambiguous red and yellow part remain problematic and require further investigation on the integratability of data (cf. Section 4). We would like to stress the need for a good technological foundation here. For example, we discovered that citypopulation.de uses Wikipedia and Wikidata as source but is likewise referenced by Wikipedia and Wikidata creating a cycle. The large size and the many details pose indeed a huge challenge, which cannot be solved by "doing more work" (cf. Section 3), but require a deeply rooted innovation.

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

Acknowledgments: This work was supported by a Wikimedia Foundation Grant GlobalFactSync[19]. We would especially like to thank the Wikipedia and Wikidata community for the huge amount of constructive feedback given during the GFS project.

References

1. Färber, M., Bartscherer, F., Menne, C., Rettinger, A.: Linked data quality of dbpedia, freebase, opencyc, wikidata, and YAGO. Semantic Web **9**(1), 77–129 (2018). https://doi.org/10.3233/SW-170275
2. Frey, J., Hofer, M., Obraczka, D., Lehmann, J., Hellmann, S.: DBpedia FlexiFusion the best of wikipedia > wikidata > your data. In: ISWC 2019. LNCS (2019). https://doi.org/10.1007/978-3-030-30796-7_7
3. Hertling, S., Paulheim, H.: Dbkwik: extracting and integrating knowledge from thousands of wikis. Knowl. Inf. Syst. **62**(6), 2169–2190 (2020). https://doi.org/10.1007/s10115-019-01415-5
4. Hofer, M., Hellmann, S., Dojchinovski, M., Frey, J.: The new dbpedia release cycle: Increasing agility and efficiency in knowledge extraction workflows. In: Semantic Systems. (2020)
5. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., et al.: Dbpedia - A large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web **6**(2), 167–195 (2015). https://doi.org/10.3233/SW-140134
6. Lewoniewski, W., Węcel, K., Abramowicz, W.: Multilingual ranking of wikipedia articles with quality and popularity assessment in different topics. Computers **8**(3), 60 (2019). https://doi.org/10.3390/computers8030060
7. Lewoniewski, W., Węcel, K., Abramowicz, W.: Modeling popularity and reliability of sources in multilingual wikipedia. Information **11**(5), 263 (2020). https://doi.org/10.3390/info11050263
8. Samuelson, Paul A.; Nordhaus, W.D.: Microeconomics (17th ed.). McGraw-Hill (2001)
9. Tanon, T.P., Vrandecic, D., Schaffert, S., Steiner, T., Pintscher, L.: From freebase to wikidata: The great migration. In: WWW (2016). https://doi.org/10.1145/2872427.2874809

[19] `https://meta.wikimedia.org/wiki/Grants:Project/DBpedia/GlobalFactSyncRE`