# Open Data Empowerment of Digital Humanities by Wikipedia/DBpedia Gamification and Crowd Curation –WiQiZi's Challenges with APIs and SPARQL

This is an extended version of the article found at https://dh2018.adho.org/revitalizing-wikipedia-dbpedia-open-data-by-gamification-sparql-and-api-experiment-for-edutainment-in-digital-humanities/

**Go Sugimoto**
ACDH-ÖAW
Vienna, Austria
go.sugimoto@oeaw.ac.at

## Abstract

Digital Humanities (DH) enjoys a wealth of Open Data published by cultural heritage institutions and academic researchers. In particular, Linked Open Data (LOD) offers an excellent opportunity to publish, share, and connect a broad array of structured data in the distributed web ecosystem. However, a real break-through in humanities research as well as its societal impact has not been visible, due to several challenging obstacles including lack of awareness, expertise, technology, and data quality. In order to remove such barriers, this article outlines an experimental case study of Application Programming Interfaces (APIs) and SPARQL. WiQiZi project employs a gamification technique to develop a simple quiz application to guess the age of a randomly selected image from Wikipedia/DBpedia. The project demonstrates a potential of gamification of Open Data not only for edutainment for the public, but also for an inspirational source of DH research. In addition, a face detection API based on an Artificial Intelligence is included for hint function, which would increase both the public and academic interests of new technology for DH. Moreover, the project provides a possibility for crowd data curation for which the users are encouraged to check and improve the data quality, when the application fails to calculate the answer. This method seems to create a win-win scenario for the Wikipedia/DBpedia community, the public, and academia.

Keywords: Digital Humanities, Application Programming Interfaces, Linked Open Data, SPARQL, gamification, crowd sourcing, data curation

## 1 Introduction

The time is ripe for Open Data. As new technology becomes available and expertise spreads across communities, governments and research entities are particularly keen to promote Open Data to meet the demands of democracy in the 21st century and ensure the transparency of research activities. In particular, Berners-Lee's (2009) five star Open Data proposition has started to take off. In his vision, Open Data is closely associated with Linked Data, which connects related data on the web with hyperlinks. He combines the two concepts and names it Linked Open Data (LOD). While the best practice of Linked Data is summarized by Heath (2018), he defines LOD as Linked Data that 'is released under an open licence, which does not impede its reuse for free'.

Followed by those initiatives, the global community has started to flourish. Over the last years, LOD has been gaining momentum in digital humanities (DH) and cultural heritage research. Many repeatedly explain the essence of LOD, including the tripod of supporting technology: HTTP, URIs, and RDF[1] (For example, Simou et al., 2017; Marden et al., 2013; Boer et al., 2016).

In fact, RDF has been adopted more frequently as a data format in data repository systems such as Fedora[2]. Important datasets including Europeana[3], VIAF[4], GeoNames[5], and Getty vocabularies[6] are being published with HTTP URIs in machine-readable formats such as RDF, Turtle[7], and JSON-LD[8]. SPARQL[9] endpoints have been progressively created in many cultural heritage organizations and DH projects (Edelstein et al., 2013). SPARQL allows the users to query a large volume of RDF graph datasets, so that semantically rich data fragments can be trans-

formed, by selecting, merging, splitting, and filtering, into new information and knowledge. 'The Promised Land' in the digital research era seems to be just around the corner.

However, research outcomes of LOD, which would have a significant impact on new discoveries and/or innovation in society, are still outstanding. Although LOD is meant to offer a powerful paradigm for global data integration, most probably reinforcing interdisciplinary research, many cases in DH are reported for the creation and publication of LOD and/or internal use of LOD (Marden et al., 2013). Although there are several DH projects concerning the use of external LOD (e.g. (Boer et al., 2016), they often focus on data enrichment. In addition, SPARQL query exploitation is rather limited within small technology-savvy communities (Lincoln, 2017; Alexiev, 2017). There could be several reasons for the underuse of LOD: a) lack of awareness of existence, b) lack of knowledge and skills to use RDF and SPARQL, c) opened data being too narrow in scope, c) lack of computing performance to be usable, and d) interdisciplinary research being not widely exercised.

In a more general framework of Open Data, the situation is much better for XML [10] and JSON[11], because they are, in general, less complex than RDF and SPARQL. As such, they are more broadly accepted as standard formats of Application Programming Interfaces [12] (APIs). However, Sugimoto ( 2017a and 2017c) is still concerned about technical hurdles for a majority of data consumers, as well as the needs of API standardization and ease of data reuse for ordinary users. In another context, the underuse of data, tools, and infrastructures seems to be a common phenomenon in DH. For example, the use of one of the most prominent services of a European language infrastructure, the Virtual Language Observatory[13] of CLARIN[14], is rather low and below expectation (Sugimoto, 2017b).

Those realities seem to indicate that research is not yet taking full advantage of Open Data, especially LOD, although a large amount of data has become available. It is a pity that the benefit of Open Data is only partially spread. To this end, this article attempts to stimulate the use of LOD within DH. The author has experimented with Wikipedia[15]/DBpedia[16] to explore the potential use of and/or the revitalization of (Linked) Open Data in and outside research community.

## 2 Gamification for Wikipedia/DBpedia (Linked) Open Data

### *2.1 Simple Quiz Application*

The choice of Wikipedia, and its structured database version, DBpedia, is rationalized by taking into account the above-mentioned issues of Open Data reuse for APIs and SPARQL endpoints. Contrary to most of DH and cultural heritage targeted projects, Wikipedia/DBpedia provides a much broader scope for data-driven research, meaning there would be more familiarity and reusability of the data among the users. This also solves the problem of datasets in DH being too specific to be used by third party researchers (or the researchers do not know how to use data and/or what to do with them (Edmond and Garnett, 2014; Orgel et al., 2015). In addition, interdisciplinary research could be more easily adopted, using a more comprehensive yet relatively detailed level of knowledge, compared to DH-branded research topics.

This paper would also serve as an example of the simple application of API and SPARQL for less technical researchers within the DH community, due to the background of this project. The project is conducted solely by the author who has developed all the code with the assistance of a colleague, albeit being a programming beginner. This setting displays an encouragement not only for researchers with less technical experience to try LOD-based research, but also for the LOD community to gain more like-minded supporters.

The project is not limited to pure research use of data. In a connection to the evolution from Open Data to Open Science (FOSTER consortium, n.d.), public interest and (ideally) engagement are just as important as the innovation potential of research itself. In this respect, the keyword of the project is **gamification**.

In order to draw public attention and to showcase a social benefit of Open Data and DH, gamification would be a catalyst to connect the scholars conducting complicated DH research and the increasingly greedy knowledge consumers among normal citizens. Kelly and Bowan (2014) states that limited attention has been paid to digital games until recently, although this is changing rapidly. Those exceptions include the recent projects of art history games reviewed by Hacker (2015). However, the intensive use of Open Data via APIs and SPARQL endpoints is still not prominent. Although there already are a few sophisticated projects such as a EU funded project Cross Cult which uses elaborate semantic

technologies (Daif et al., 2017), this article is able to contribute to this discourse from a web innovation perspective in a more simplified DIY project environment.

The primary outcome of the project is WiQiZi[17], a simple quiz application, based purely on external Open Data APIs and SPARQL. In a nutshell, it requires users to guess the age of a randomly selected person from Wikipedia by looking at a portrait of the person.

The game starts with a selection of a year in order to specify the time of the target person (Fig. 1). Ten random years between 1700 and 2002 are generated and presented to the users. It is recommended to pick one of them, because the year range is more likely to find a person from a pool of available people. The users can also type a specific year of their choice. The year is used as the birth year of the person. When an image of a person is loaded, the users can start guessing the age of the depicted person, also using the description of the person as a clue (Fig. 2 and Fig. 3). As such, WiQiZi represents an interplay of **W**ikipedia, **Q**ui**Z** and **I**nformation, delivering elements of entertainment, education and research for potentially a wide range of audience.

Apparently, the age of a person in a particular image is provided neither by Wikipedia, nor by DBpedia. It is, in fact, calculated programmatically by comparing the birthdate and the creation date of the image. Although it is a simple algorithm, the quiz is generated automatically. It goes without saying that this approach does not guarantee the correct answer. For example, an image may be created after the death of a person. If the image is a photograph, it is likely to be more accurate. Thus, this game merely provides the best guess based on available facts. Nevertheless, it is good enough for edutainment, because the main purpose of the application is to stimulate the users' interest. In addition, it only takes into account years but not months or days. On a positive side, the application enables users to play the game even if either (or both) of months and days are missing (see Section 3 too for data quality issues).

The random selection of data is sometimes costly for data processing, but it was applied for year and image in the application. Randomization is, in fact, the key to developing a game application, as gamers easily get bored, if the game always shows the same information and situation. The application is intended for fun, thus, includes both female and male, and all types of contemporary persons such as politicians, sport athletes, musicians, actors, and businesspersons. Living persons are useful to increase the engagement level of the users. At the same time, the inclusion of historical figures is very important in DH in that the user would learn the history of a person from the past. As a result, figures range from Oliver Cromwell (political leader) and Luis Peglion (bicycle racer) to Irina Shayk (model) and Ariana Grande (singer). In this regard, the project successfully represents the richness and diversity of information which LOD can offer for history, art and culture, media studies, and alike.

The images are typically paintings, drawings, prints, photos, but occasionally objects such as statues and coins, which depict a person. It is also possible that no person is depicted in the image. For instance, they can be graves or items that symbolise the person (Fig. 4). The earlier the year, the more likely it is that the image does not contain the portrait of a person. In such cases, users are required to reshuffle the image (see yellow box in Fig. 2).

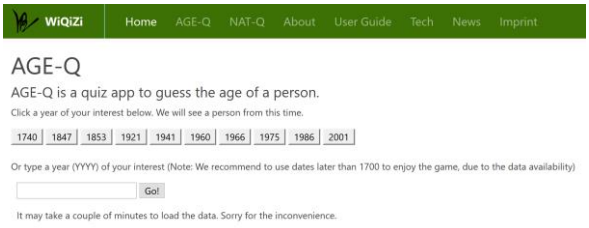


Fig. 1 Select a randomly generated year, or type a year in the text box

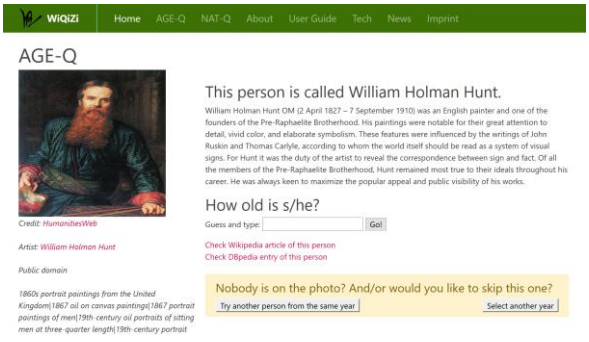


Fig. 2 Quiz to guess the age of a person found in a Wikipedia article

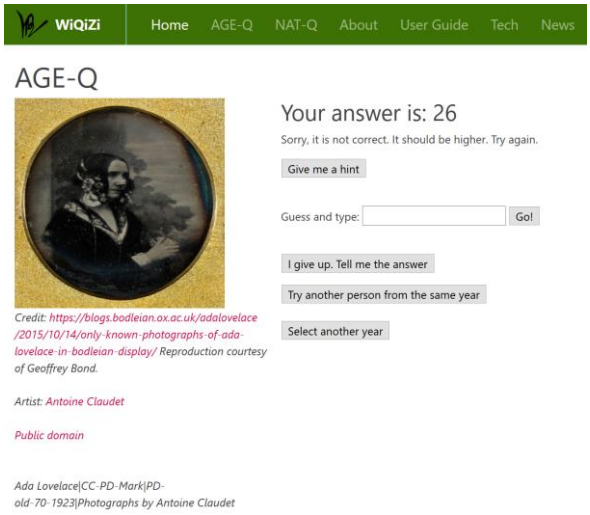


Fig. 3 The screen when submitting a wrong answer

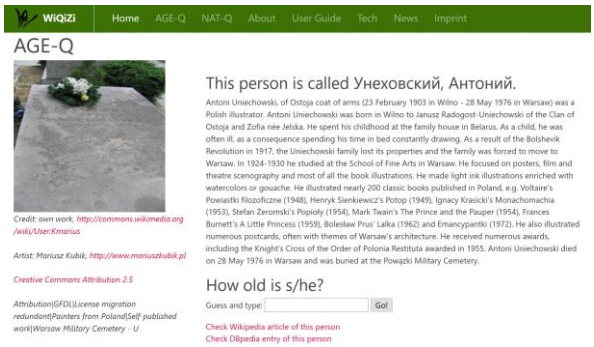


Fig. 4 Symbol of a person in the image

When the user cannot guess the age, there is a help function. A hint section is equipped with a face detection API, suggesting the estimate age and gender of the person in the image by machine learning (Fig. 5). The confidence score of the estimation is also given by percentage. Although the function is extremely simple, the current boom of Artificial Intelligence in our society would inspire DH and alike in the context of APIs in combination with Open Data from Wikipedia/DBpedia. When the right answer is delivered, the application displays the links to the corresponding Wikipedia article and DBpedia dataset. This gives the users opportunity to learn the person in detail.

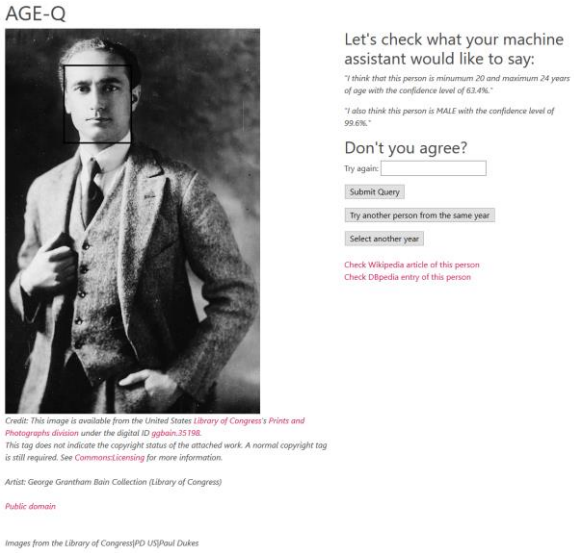


Fig. 5 Hint function for age and gender, using the face detection and machine learning API

## *2.2 Simple Technology, but Technical Challenges*

The application is built with simple PHP[18] even without using a framework such as Laravel[19]. Bootstrap[20] is used for creating a quick web design. This is why the project would be better labelled as a DIY project. The application is entirely based on external data via APIs and SPARQL endpoint, exploring the potential of distributed data research. It uses three different APIs of Wikipedia[21] and SPARQL endpoints of DBpedia[22]. The former consists of 1) Wikipedia API to access Wikpedia articles, 2) Wikimedia API to access images in Wikipedia articles, and 3) another Wikimedia API to access the metadata of the images. The third one is crucial in that it contains copyright and licence information. All the images of the quiz come with as much IPR information as possible, so that the application ensures the protection of copyright, while promoting data re-use by clarifying the licenses.

It should be also noted that WiQiZi does not store any images on the server. It displays images directly from Wikimedia, being a lightweight software application. It is worth iterating that the automatic generation of quiz is neither very common nor easy, because a quiz has to provide intellectual challenges and the right level of difficulty, thus, many quizzes are hand-written. Thanks to the semantics of DBpedia, question-answer applications such as WiQiZi can be developed.

The mix of APIs makes the application development a little tricky, because API calls have to

be made one after another, although they all serve the data that originate, one way or another, from Wikipedia. Better organization of data access to those resources would increase the usability of the developers and users in the future. In contrast, there are also advantages for the use of APIs in a decentralized system. It allows developers to save resources (cost of servers and storage and maintenance) and to focus on system and data integration.

For the sake of data pursing, a SPARQL query is embedded into API query parameters and the query results are returned as JSON[23]. DBpedia automatically executes this transformation. For example, the following SPARQL query (whose results can be seen with the DBpedia endpoint interface on a web browser: Fig. 6) can be transformed into API query parameters below:

```
SELECT *
WHERE {?person rdfs:label ?person_name
; rdf:type ?type ; dbo:birthDate
?birthdate ; dbo:abstract ?abstract .
    bind(rand(1 +
strlen(str(?person))*0) as ?rid)
FILTER regex(?type, "Person")
FILTER regex(?birthdate, "1977")
} order by ?rid
LIMIT 200
```

```
http://dbpedia.org/sparql?default-
graph-
uri=http%3A%2F%2Fdbpedia.org&query=sele
ct+*%0D%0Awhere+%7B%3Fperson+rdfs%3Alab
el+%3Fperson_name+%3B+rdf%3Atype+%3Ftyp
e+%3B+dbo%3AbirthDate+%3Fbirthdate+%3B+
dbo%3Aabstract+%3Fabstract+.%0D%0A++++b
ind%28rand%281+%2B+strlen%28str%28%3Fpe
rson%29%29*0%29+as+%3Frid%29%0D%0AFILTE
R+regex%28%3Ftype%2C+%22Person%22%29%0D
%0AFILTER+regex%28%3Fbirthdate%2C+%2219
77%22%29%0D%0A%7D+order+by+%3Frid%0D%0A
LIM-
IT+200&format=json&CXML_redir_for_subjs
=121&CXML_redir_for_hrefs=&timeout=3000
0&debug=on&run=+Run+Query+
```

Fig. 6 SPARQL query results (part)

The implementation of face detection is also simple. The application posts an image of the quiz as URL to the image analysis API of IBM Watson[24]. The API returns JSON data with the estimation of the age and gender of the person depicted, as well as the numeric location of the facial area. If the image is larger than the standard layout of the game interface, it should be adjusted accordingly. In that case, extra PHP coding is needed to calibrate the area of the face by using the ratio of resize.

There are a couple of technical challenges. First of all, it turns out that the DBpedia dataset is not as rich as one may expect. More precisely, if a SPARQL query is fired to access a generic dataset, for example, to select data classified as 'person', there are often only few shared RDF properties in the query results (Table 1). Namely, name of the person, description, and link to the Wikipedia article. Even birthdates and birthplaces (and death date and place) may not exist, depending on the data quality.

In addition, the occupation of the person determines the availability of his/her properties. For instance, whereas football players may have properties related to club teams and national caps, politicians hold properties related to political parties and experience of ministers, etc. This generalization-specialization makes it hard to anticipate what properties are available for different persons. This is not a problem for DBpedia; however, it is a challenge for a quiz application, which has to start with a generic query in order not to preselect the DBpedia categories of persons.

In fact, the very first SPARQL query of AGE-Q is to randomly retrieve data from entries of the type "person" that have user-selected year for the variable of birthdate (See above and Fig. 6.). A further condition is set in PHP to restrict the data to ones with thumbnails available. Unless an alternative interface is developed (e.g. select occupation first) and the quiz compromises the amount of available persons, the quiz questions need to be very generic. This is the very reason why age was chosen for the application in the first place.

Table 1 Summary of available RDF property

| Available level | Likely available common-properties |
|---|---|
| Almost always | Name (rdf:label), type (rdf:type) description (dbo:abstract), Wikipedia link[25] |
| Frequent | Birthdate (dbo:birthDate), birthplace, death date, death place, nationality etc. |
| Sometimes | Spouses, occupation, associated people etc. |
| Depending on the type of person | Art works, publications, political parties, teams, etc. |

Secondly, although rather trivial, the application currently does not support face detection for multiple persons in an image. Therefore, it may not return the estimation of the right person. In rare cases, there are a multiple persons in an image and one of them is the very person of the Wikipedia article. At the moment, there is no excellent logic to identify the face of a person in question, and filter out the others. IBM Watson

normally detects several faces without prioritising them.

Thirdly, as hinted earlier, the performance is slow, when loading the first image. In the worst case, it might take up to a couple of minutes to load the quiz, because the application depends on the chain of APIs and select data randomly. Although the users are informed on the start page, a progress bar is not yet implemented. In the long run, the code needs to be refactored and optimized in order to satisfy the users. As Sugimoto (2017a) reported, the chain of API calls opens an avenue for a new data mash-up possibility, but the current web technology may not be sufficient for pragmatic use cases of such distributed systems.

Lastly, the application has no multilingual support. The description of a person is always in English, while person names (used as a header of the quiz page) may be presented in languages other than English (See Fig. 3). The biggest obstacle of the multilingual extension of WiQiZi is SPARQL query. It is assumed that swapping language code (e.g. xml:lang="en" to xml:lang="ja") is enough to convert English game to Japanese one. However, it turns out that it is not possible to re-use the SPARQL query used for English DBpedia for another language version of DBpedia, because each language version of DBpedia uses a different ontology. Only a fraction of the ontology (such as rdfs:label and rdf:type) is the same across different languages. For example, the RDF property, http://dbpedia.org/ontology/birthDate, is replaced by http://es.dbpedia.org/property/nacimiento or http://es.dbpedia.org/property/fechaDeNacimiento ("fecha de nacimiento" is the Spanish translation of birthdate) for Spanish DBpedia. While Dutch DBpedia uses dbpedia-owl:birthdate instead, Italian DBpedia has another property called http://it.dbpedia.org/property/annonascita as well as dbpedia-owl:birthdate.

Many of the variations of property names are unpredictable. This makes it complicated to replicate the game in the same manner. There is also inconsistency between different languages of DBpedia, causing confusion for data integrity. The data organization problem between Wikipedia and DBpedia only adds complications to the data quality discussion. Therefore, we must acknowledge that although DBpedia provides extremely useful structured data, it has not yet become a fully reliable source of information for serious research.

# 3 Ongoing Development and Future Work

## *3.1 Potential of Gamification and Citizen Science*

Another use of this application for the empowerment of DH and Wikipedia communities is the crowdsourcing of the curation of Wikipedia articles and DBpedia datasets. Data curation is one of the burgeoning issues of DH and cultural heritage. Countless publications are produced every year to discuss the data quality in the field of library science, archives and DH in general. For instance, as a reflection of critics of Linked Data quality, Daif et al. (2017) reckon that human supervision is needed to manage the data in their project.

In our case, the application is sometimes not able to calculate the age of a person, due to several reasons of metadata quality. For instance, data may be not numeric (e.g. "16th century") (See Fig. 7 for a Wikipedia/Wikimedia case), malformed (e.g. not ISO compliant: "05/11/88"), confusing (e.g. the creation date of digital image is used instead of that of analogue image), inaccurate (e.g. 1880s instead of 1885 (true value) due to uncertainty), wrong (e.g. 2599 instead of 1599 due to mistype), or missing, resulting in an error message.

This is normally regarded as an optimization problem of the code. While usually developers might try to suppress erroneous results, in this case, we are not interested in concealing errors. When the error occurs, it could be a sign of a data quality problem and we could trace back to underlying inconsistencies in the data structure. In this application, users are persuaded to follow the provided links to Wikipedia and DBpedia and able to double-check the original data (Fig. 8). If the users are able to correct and/or improve data, for instance, by executing a little online research, the impact for data curation could be considerable. This scenario creates a dual possibility. In other words, the application can be used as:

- A curation tool of Wikipedia and DBpedia for existing active editors of Wikipedia.
- A tool to transform normal users into new curators of Wikipedia

Not only could crowd curation benefit Wikipedia by correcting and/or adding data, but DBpedia would also be improved, leading to a higher quality of datasets of this LOD magnet

and affecting hundreds of applications worldwide.


Fig. 7 Wikimedia metadata displaying non-numeric date ("16th century")

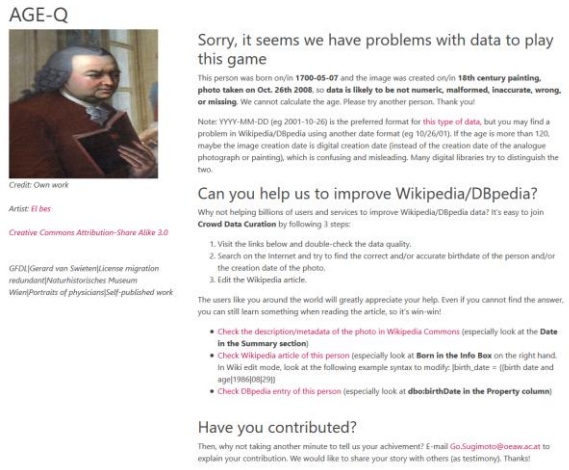


Fig. 8 Crowd sourcing potential for the game application

Without doubt, crowdsourcing has become an established subject in DH in its own right. Wikipedia itself is an exemplar of crowdsourcing (Carletti et al., 2015). In her introductory article about crowdsourcing in DH, Terras (2016) reported and discussed an overview of crowdsourcing examples applied in DH and the GLAM (Galleries, Libraries, Archives and Museums) sector, ranging from an early example of the Australian Digitisation Program[26] (Holley, 2009) and the North American Bird Phenology Program[27] to Transcribe Bentham and Soldier Studies[28].

With regard to data curation, the success of crowdsourcing is proven by examples such as "Wasisda?" by the Sound and Vision in the Netherlands (Brinkerink, 2010) , and "What's on the menu?" by the New York Public Library (NYPL Labs). Focusing on user engagement and its design patterns, Ridge, (2013) analysed the success of Old Weather[29], Herbaria@Home[30], and Galaxy Zoo[31].

Dunn and Hedges (2012) states that one of the four factors of crowdsourcing used within humanities research is a clearly defined core research question and direction within the humanities. This argument is also echoed in other literatures (Ridge, 2013; Terras, 2016). In this regard, however, WiQiZi project has a different view. It does not define a clear research question for the crowdsourcing; thus it does not help a specific area of humanities research per se. Rather, it brings a humanities and/or cultural heritage perspective of and inspiration for a new use of Wikipedia/DBpedia for scholars.

In addition to that, the author prefers to focus on the gamification of 'reasonably intellectual' materials (i.e. Wikipedia), and the crowd sourcing possibility is regarded as a spin-off service. The advantage of this positioning is that it emphasizes on the 'voluntary' public engagement of Wikipedia data curation, rather than the 'mission' of institutional crowdsourcing that often has a certain goal, expectation, or ambition to complete relatively specific tasks. A disadvantage is the lower level of public participation. The more voluntary and supplementary the crowdsourcing becomes, the less the users would engage and help.

In this respect, WiQiZi takes a slightly unusual approach to crowdsourcing. Therefore, it may not be covered by the crowdsourcing typology (See for example Carletti et al. (2015)). Moreover, integrating gamification and crowdsourcing would be an answer for our project to enhance the motivation of the participants. WiQiZi implements a kind of crowdsourcing possibility in such a way that the participants join it without noticing or are less conscious about it.

As Ridge (2013) observed, crowdsourcing participants can be categorised into two: those who are intentionally participating and those whose contributions are a side effect of their participation in other core activities. Cases where the core activity of the latter is a game may be called crowdsourcing games, which seems to better fit the classification of WiQiZi. Subsequently, WiQiZi project is comfortably in sync with the DH advocators who are careful about criticism on a potential risk of labour exploitation (Terras, 2016).

Currently there is no good mechanism implemented to systematically collect information about the data curation. The users are encouraged to contact the developer via email to report their contributions to data curation. However, adding such a complication affects the incentive to play

the game and constitutes extra effort. It could be possible to track user engagement in Wikipedia, using user logs. Unfortunately, this would require a good deal of elaboration to the application, therefore it is not planned in the close future.

Concerning the combat for data quality on the web, there are initiatives such as a W3C working group, which works on creating a Data Quality Vocabulary (DQV)(W3C, 2016). It will not specify what quality means, simply because some datasets are useful for some, but low quality for other purposes. Instead, it aims to make it easier to publish, exchange, and consume quality metadata for every step of a dataset's lifecycle. In this way, different stakeholders can evaluate the datasets and the data consumers are free to use it as an aid to assess data quality by themselves.

The author understands that the need of such a vocabulary has arisen from the situation where it has become difficult to find valuable datasets in the enormous sea of data on the web. In the future, metadata such as DQV could help the users to identify the data quality including DBpedia.

### *3.2 Future Work*

Improvement to the application could be made on several levels. For example, it would be very interesting to have a point awarding system. Incentivization is arguably one of the most challenging parts of crowd sourcing, as Ridge (2013) explored the user motivation and engagement.

One simple addition would be to display the amount of guessing attempts to reach the right answer. Based on the count, points can be given to the users, adding more fun element to the game. By introducing a login registration, points can be saved to the user account. There is no doubt that the point system is effective, especially when the game stimulates user competition.

In addition, an even more ambitious system can be developed. Ideally, more points should be awarded for contributions to the crowd data curation. Although it is not an easy task to consider a fair way of validation and score provision, for example, by assigning moderators in the user community, it would surely increase user engagement. For this type of deep engagement, Ridge (2013) suggests the use of scaffolding techniques of museums for online crowdsourcing. Scaffolding design provides clear user roles and information about participation. It also carefully manages the complexity level of participation with a shallow learning curve and guidance through early levels of participatory activities.

Obviously, it is not trivial to devise a sophisticated platform that deploys such scaffolding, due to the DIY nature of WiQiZi. However, if more interactive features such as score comparison, personalisation, and visibility of contribution are implemented, it would be a win-win situation for users to enjoy the game and the Wikipedia/DBpedia community to gain more voluntary support in terms of data curation.

In this project, the official LOD version of Wikipedia, Wikidata[32], has not been explored. The tight connection between Wikipedia and Wikidata would provide an outstanding chance for WiQiZi. Depending on the data quality of Wikidata, WiQiZi can be extended to more detailed quiz questions, which will increase the appetite of the users.

On the other hand, the use of Wikipedia/DBpedia was just a beginning of effective LOD research in DH. Its full potential can be examined only by stretching the data integration to other data sources. To this end, Europeana is in the scope of the next development, which supplies over 50 million cultural heritage objects of Europe. Its metadata is offered with CC0 license.

Good linking points between Europeana and Wikipedia/DBpedia, as well as other such important resources as GeoNames, VIAF, and Getty Vocabularies need to be investigated, so that the applications like WiQiZi would truly incentivize DH research based on the fully-fledged LOD cloud.

Multilingual support is also needed for the promotion of data diversity. It is the interest of not only the DH community in which language plays a vital part of research, but also the Wikipedia/DBpedia community, as well as the web community at large, which facilitates diversity. As seen in Section 2, like many other web projects, Wikipedia and DBpedia are rather unnaturally English-oriented. Wikipedia's multilingual achievement is extraordinary in the sense of local community development.

In contrast, DBpedia seems to be lagging behind. For many LOD experts, the English version takes the central position, partly because of the richness of data. The development of language chapters is rather slow. While there are 292 active Wikipedia language versions (Wikipedia, 2018), only about 20 DBpedia versions exist (DBpedia, 2018)(Fig. 9).

Many DBpedia websites are operated on a voluntary basis by local communities that lack

organisational, technical, and financial support. According to a survey among nine DBpedia language chapters, 66% of the chapters have only one to four people involved in the core work, while only one chapter has about ten people (DBpedia Association, 2018a). In addition, 44.4% update their services once a year, while over 22.2% have not updated in more than two years (DBpedia Association, 2018b). Furthermore, the inconsistencies of their websites and the lack of cooperate design reflect the backlog of the multilingual versions of the project. If WiQiZi were able to cope with different language versions, it would help to promote DBpedia chapters and could be presented as a use case for multilingual LOD.

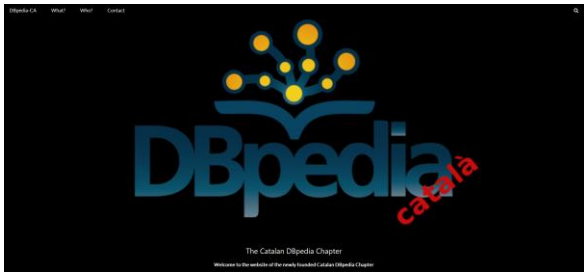


Fig. 9 Catalan DBpedia chapter

## 4 Conclusion

In conclusion, this article demonstrates an experimental case study of mixing gamification techniques (entertainment) with data-driven research (education) and the possibility for data curation (crowdsourcing), showcasing cutting-edge technologies such as SPARQL and Deep Learning API, with the help of Open Data in the framework of DH. In addition, it presents an example of an application that automatically generates simple quizzes, based on semantic question-answer capability. Moreover, it displays a potential for a new digital research ecosystem for humanities research and digital technologies, connecting various stakeholders including humanities researchers and the public.

As Terras, (2016) puts it, 'the Digital Humanities can aid in creating stronger links with the public and humanities research, which, in turn, means that crowdsourcing becomes a method of advocacy for the importance of humanities scholarship, involving and integrating non-academic sectors of society into areas of humanistic endeavour.'

It would be interesting to use the same or a similar method for different types of quiz. Given the variety and richness of information in Wikipedia, the automatized quiz can be about any interesting concept including buildings, objects, places, events, genres, and movements. Hence, Wikipedia seems to provide an exciting platform for edutainment, especially in the art and humanities sphere. In addition, the project plans to continue developing a more elaborate game application by taking advantage of semantically rich Open Data resources such as Europeana, VIAF, GeoNames, and Getty Vocabularies. If WiQiZi could cope with DBpedia multilingual chapters, it would be able to become a valuable prototype for the representation of LOD diversity.

At the same time, the paper also acknowledges several challenges. For instance, technical developments such as improvement of query performance are required in order to use LOD for practical day-to-day research business. Standardization may be required to lower the barrier of the complex technical environments especially for the ordinary yet majority of users. There are also data quality issues for Wikipedia and DBpedia to be fully useful for serious research, especially in the context of automatization. Raising awareness is another social issue.

Admittedly, although the application of this project is fairly simple, it is hoped that it helps to inspire and incentivize the DH researchers to actively use LOD as a new tool for our knowledge society in which any member of the society can become an active actor of knowledge creation, curation, and distribution.

## Notes

[1] https://www.w3.org/RDF/
[2] https://fedorarepository.org/
[3] https://pro.europeana.eu/resources/apis/sparql
[4] https://viaf.org/
[5] http://www.geonames.org/
[6] http://vocab.getty.edu/
[7] https://www.w3.org/TR/turtle/
[8] https://json-ld.org/
[9] https://www.w3.org/TR/sparql11-overview/
[10] https://www.w3.org/XML/
[11] https://www.json.org/
[12] https://en.wikipedia.org/wiki/Application_programming_interface
[13] https://vlo.clarin.eu
[14] https://www.clarin.eu/
[15] https://www.wikipedia.org/
[16] http://wiki.dbpedia.org/
[17] https://wiqizi.acdh-dev.oeaw.ac.at
[18] http://www.php.net/
[19] https://laravel.com/
[20] https://getbootstrap.com/

[21] https://www.mediawiki.org/wiki/API:Main_page
[22] http://dbpedia.org/sparql
[23] http://www.json.org/
[24] https://www.ibm.com/watson/services/visual-recognition/
[25] The Wikipedia URL is easily inferred by the corresponding slug URL of DBpedia. The application excluded foaf:primaryTopic to fetch Wikipedia link (partly in order to increase query performance).
[26] https://www.nla.gov.au/content/newspaper-digitisation-program
[27] http://www.birds.cornell.edu/citscitoolkit/projects/pwrc/nabirdphenologyprogram/
[28] http://www.soldierstudies.org/
[29] https://www.oldweather.org/
[30] http://herbariaunited.org/atHome/
[31] https://www.zooniverse.org/projects/zookeeper/galaxy-zoo/
[32] https://www.wikidata.org

## References


**Alexiev, V.** (2017). Getty Vocabularies LOD: Sample Queries http://vocab.getty.edu/queries#Finding_Subjects (accessed 2 October 2018).

**Berners-Lee, T.** (2009). Linked Data - Design Issues https://www.w3.org/DesignIssues/LinkedData.html (accessed 2 October 2018).

**Boer, V. de, Penuela, A. M. and Ockeloen, C. J.** (2016). Linked Data for Digital History: Lessons Learned from Three Case Studies. *Anejos de La Revista de Historiografía*(4): 139–62.

**Brinkerink, M.** (2010). Waisda? Video Labeling Game: Evaluation Report *Images for the Future – Research Blog* http://research.imagesforthefuture.org/index.php/waisda-video-labeling-game-evaluation-report/index.html (accessed 2 October 2018).

**Carletti, L., Giannachi, G., Price, D., McAuley, D. and Benford, S.** (2015). Digital humanities and crowdsourcing: an exploration. https://core.ac.uk/reader/43093962 (accessed 2 October 2018).

**Daif, A., Dahroug, A., López-Nores, M., Gil-Solla, A., Ramos-Cabrer, M., Pazos-Arias, J. J. and Blanco-Fernández, Y.** (2017). Developing Quiz Games Linked to Networks of Semantic Connections Among Cultural Venues. *Metadata and Semantic Research*. (Communications in Computer and Information Science). Springer, Cham, pp. 239–46 doi:10.1007/978-3-319-70863-8_23. https://link.springer.com/chapter/10.1007/978-3-319-70863-8_23 (accessed 2 October 2018).

**DBpedia** (2018). Chapters. https://wiki.dbpedia.org/join/chapters (accessed 2 October 2018).

**DBpedia Association** (2018a). DBpedia Chapters – Survey Evaluation – Episode One | DBpedia https://wiki.dbpedia.org/blog/dbpedia-chapters-%E2%80%93-survey-evaluation-%E2%80%93-episode-one (accessed 2 October 2018).

**DBpedia Association** (2018b). DBpedia Chapters – Survey Evaluation – Episode Two | DBpedia https://wiki.dbpedia.org/blog/dbpedia-chapters-%E2%80%93-survey-evaluation-%E2%80%93-episode-two (accessed 2 October 2018).

**Dunn, S. and Hedges, M.** (2012). Crowd-Sourcing Scoping Study Engaging the Crowd with Humanities Research.

**Edelstein, J., Galla, L., Li-Madeo, C., Marden, J., Rhonemus, A. and Whysel, N.** (2013). Linked Open Data for Cultural Heritage: Evolution of an Information Technology. http://www.whysel.com/papers/LIS670-Linked-Open-Data-for-Cultural-Heritage.pdf (accessed 2 October 2018).

**Edmond, J. and Garnett, V.** (2014). Building an API is not enough! Investigating Reuse of Cultural Heritage Data *LSE Impact Blog* http://blogs.lse.ac.uk/impactofsocialsciences/2014/09/08/investigating-reuse-of-cultural-heritage-data-europeana/ (accessed 2 October 2018).

**FOSTER consortium** (n.d.). What is Open Science? Introduction, *FOSTER FACILITATE OPEN SCIENCE TRAINING FOR EUROPEAN RESEARCH* https://www.fosteropenscience.eu/content/what-open-science-introduction (accessed 2 October 2018).

**Hacker, P.** (2015). The Games Art Historians Play: Online Game-based Learning in Art History and Museum Contexts *The Chronicle of Higher Education Blogs: ProfHacker* https://www.chronicle.com/blogs/profhacker/the-games-art-historians-play-online-game-based-learning-in-art-history-and-museum-contexts/61263 (accessed 12 April 2018).

**Heath, T.** (2018). Linked Data | Linked Data - Connect Distributed Data across the Web http://linkeddata.org/home (accessed 2 October 2018).

**Holley, R.** (2009). A success story - Australian Newspapers Digitisation Program Journal article (Paginated) *Online Currents* http://eprints.rclis.org/14176/ (accessed 2 October 2018).

**Kelly, L. and Bowan, A.** (2014). Gamifying the museum: Educational games for learning | MWA2014: Museums and the Web Asia 2014 https://mwa2014.museumsandtheweb.com/paper/gamifying-the-museum-educational-games-for-learning/ (accessed 2 October 2018).

**Lincoln, M.** (2017). Using SPARQL to access Linked Open Data. *Programming Historian* https://programminghistorian.org/lessons/graph-databases-and-SPARQL (accessed 2 October 2018).

**Marden, J., Li-Madeo, C., Whysel, N. Y. and Edelstein, J.** (2013). Linked Open Data for Cultural Heritage: Evolution of an Information Technology. *Columbia University Academic Commons* https://doi.org/10.7916/D89021QD (accessed 2 October 2018).

**NYPL Labs** Whats on the menu? http://menus.nypl.org/about (accessed 2 October 2018).

**Orgel, T., Höffernig, M., Bailer, W. and Russegger, S.** (2015). A metadata model and mapping approach for facilitating access to heterogeneous cultural heritage assets. *International Journal on Digital Libraries*, **15**(2–4): 189–207 doi:10.1007/s00799-015-0138-2.

**Ridge, M.** (2013). From tagging to theorizing: deepening engagement with cultural heritage through crowdsourcing. https://core.ac.uk/display/82977685?source=2&algorithmId=14&similarToDoc=82981870&similarToDocKey=CORE&recSetID=8d9b82c8-62e3-430e-91e7-ee235bcff1ea&position=3&recommendation_type=same_repo&otherRecs=19758508,41340087,82977685,43093962,147827629 (accessed 2 October 2018).

**Simou, N., Chortaras, A., Stamou, G. and Kollias, S.** (2017). Enriching and publishing cultural heritage as linked open data. In Ioannides, M., Magenat-Thalman, N. and Papagiannakis, G. (eds), *Mixed Reality and Gamification for Cultural Heritage*. Springer http://eprints.lincoln.ac.uk/26895/ (accessed 2 October 2018).

**Sugimoto, G.** (2017a). Who is open data for and why could it be hard to use it in the digital humanities? Federated application programming interfaces for interdisciplinary research. *International Journal of Metadata, Semantics and Ontologies*, **12**(4): 204 doi:10.1504/IJMSO.2017.10014806.

**Sugimoto, G.** (2017b). Number game -Experience of a European research infrastructure (CLARIN) for the analysis of web traffic. *CLARIN Annual Conference 2016*. Aix-en-Provence, France: CLARIN ERIC and Laboratoire Parole et Langage and Laboratoire des Sciences de l'Information et des Systèmes (LSIS) and Aix-Marseille Université and Centre National de la Recherche Scientifique (CNRS) https://hal.archives-ouvertes.fr/hal-01539048 (accessed 2 October 2018).

**Sugimoto, G.** (2017c). Battle Without FAIR and Easy Data in Digital Humanities. *Metadata and Semantic Research*. (Communications in Computer and Information Science). Springer, Cham, pp. 315–26 doi:10.1007/978-3-319-70863-8_30. https://link.springer.com/chapter/10.1007/978-3-319-70863-8_30 (accessed 25 April 2018).

**Terras, M.** (2016). Crowdsourcing in the Digital Humanities. *A New Companion to Digital Humanities*. Wiley-Blackwell, pp. 420–439 https://hcommons.org/deposits/download/hc:15066/CONTENT/

mterras_crowdsourcing20in20digital20humanities_final1.pdf/ (accessed 2 October 2018).

**W3C** (2016). Data on the Web Best Practices: Data Quality Vocabulary https://www.w3.org/TR/vocab-dqv/ (accessed 2 October 2018).

**Wikipedia** (2018). List of Wikipedias. https://meta.wikimedia.org/wiki/List_of_Wikipedias (accessed 2 October 2018).