Published in proceedings of the International Conference on Advanced Computer Science and Information Systems (ICACSIS) 2017, Bali, Indonesia, 28-29 October 2017, which should be cited to refer to this work. DOI:10.1109/ICACSIS.2017.8355013

Enhancing Cultural Heritage Data through Selective Crowdsourcing

Alex C. Olivieri *, Shaban Shabani †, Zhan Liu ‡, Maria Sokhn §
Data Semantics Lab,
Institute of Information Systems,
HES-SO Valais-Wallis,
Sierre, Switzerland
Email: *alex.olivieri@hevs.ch, †shaban.shabani@hevs.ch, ‡zhan.liu@hevs.ch, §maria.sokhn@hevs.ch

***Abstract*—Crowdsourcing, or asking for general participation to help contribute to shared goals, has become popular in the cultural heritage domain. However, it also brings new challenges to data quality management. In this paper, we describe a novel approach that used selective crowdsourcing to increase quality control and the number of participants when collecting cultural heritage-related data. We tested our approach in scenarios using a mobile application for cultural heritage travellers. The results showed that the selective crowdsourcing approach can be applied well to complete information and resolve conflict in cultural heritage data.**

I. Introduction

Cultural heritage is something people should try to preserve because it connects us to our roots. It is transmitted orally from generation to generation when it involves intangible material, or by inheritance when it refers to physical artifacts [1]. Even if museums and foundations help diffuse such knowledge, cultural heritage represents a peculiar domain and those who are inexpert can experience limitations when searching for information. The reason is that the content, even if available, is distributed among various libraries, archives, and institutions, and it is challenging to integrate the information available and obtain an aggregate view that is useful to a common audience.

This represents a shortcoming for many domains in which cultural heritage is a topic of interest, especially in the tourism industry, where more and more tourists are changing the way they engage in tourism. As short trips to visit nearby locations with significant cultural heritage appeal is overtaking the habitual long trips tourist agencies often organize [2] [3] [4], it has become crucial to integrate such information. Moreover, we believe that creating an application that helps tourists discover, explore, and visit cultural heritage sites by providing access to this aggregated information can have a considerable influence on this new form of tourism.

However, because cultural heritage is transmitted over generations, it can suffer from missing information. Moreover, because the data are distributed in various institutions, they can include conflicting content that provides misleading information. To address these issues, we believe that a joint contribution on the part of tourism experts and tourists can improve the quality of the information available, because it allows us to merge the specific knowledge provided by experts in tourism with the large amount of data that can be provided by crowd participants.

In this paper, we present a mobile application that uses aggregated cultural heritage data provided by the Cityzen platform. This application helps tourists organize visits to cultural heritage locations. Advances in mobile computing have led to mobile crowdsourcing [5], a new approach to enhancing data collection and processing. We used mobile crowdsourcing as a service that enriches the data repository with new records and at the same time, corrects existing data. The application allows crowd participants to share their data related to the history of their country, thereby contributing to the quality of the data. Given that data about cultural heritage must be accurate, their quality is very important. Therefore, the mobile application was used to ask participants to improve the quality of the data by correcting errors that appear in the repository. Moreover, we applied the selective crowdsourcing approach, a mechanism used to filter crowd participants to maintain the quality of crowdsourced data.

The paper is organized as follows: in section II, we describe the state of the art in the domain, and highlight its limitations; in section III, we describe our previous work on the Cityzen platform; in section IV, we introduce our application to support tourists in planning and engaging in sightseeing, and the built-in feature for users participation; in section V, we explain the selective crowdsourcing methodology used to improve the information content actively, and in section VI, we conclude the paper by discussing the advantages of our work and offering suggestions for future research.

II. Related Work

Researchers have experimented for many years with the convergence of internet and wireless technologies to create web platforms to publish information of cultural interest that allows visitors to exploit cultural heritage material before, during, and after their visits. [22] focused on the collection and modelling of users behaviors, including their interests, personal characteristics, and contextual factors, to build a personalization system in cultural heritage. The findings presented a way to match visitors and cultural heritage-based activities. In [23]'s study, the authors introduced an OCR processing and transcription method to create an online application for

the complete management of cultural heritage information. In their results, they proposed a new interoperability and standard structure that could be used to guarantee the usability and usefulness of cultural heritage records. [24] used metadata schemas and controlled vocabularies to collect and complete cultural heritage information, and the methodology ensured the processes of metadata schema selection and good end-user access. A recent study [25] presented systems integration for information collection and analysis of heterogeneous cultural heritage data in museums. They used a case study to explain the way in which to integrate existing information systems to improve museums operational performance in a climate of growing competition and increasing complexity. These solutions provide different technologies to collect cultural heritage information. However, the failure to consider conflicting and missing information could have biased the research results.

We used Cityzen [6] to create an application for cultural heritage tourism that provides information for people who take cultural heritage holidays. This application is based on a data model designed for tourism purposes, and is populated by aggregated data provided by various institutions that own data that contain cultural heritage information. As we noticed while developing this application, the data present problems that we divided into two different categories: conflicting information [7], and missing information [8],[9]. In previous work, we dealt with such problems by asking users for their contributions, as suggested in [11]. This approach led to some qualitative improvement in the data because users were asked to correct missing or conflicting information associated with their interests, and usually were motivated and had knowledge to share. However, the number of contributions was limited because the application has not had a large number of users in the past.

Therefore, to address the issue of participation, we used crowdsourcing and invited people through online channels, such as social media and websites, to use the application and participate in data annotation. Crowdsourcing has been applied widely to solve data-related problems, especially those that are difficult for computers, but trivial for humans. Such applications include the entity resolution field [14], sentiment analysis [15], and image classification [16], among others.

Crowdsourcing has the potential to create a more open, connected, and smart cultural heritage [17] by involving users, consumers, and providers. However, it is difficult and challenging to find users who will provide valuable information, and hence maintaining the quality of the data provided is a major issue in crowdsourcing.

Different techniques have been proposed to address quality control. A basic technique that online crowdsourcing platforms have applied already is the qualification test, which is designed to eliminate unskilled participants by asking questions to which the answers are known in advance [20]. Other mechanisms [19] are redundancy-based techniques that ask several participants about the same task, and by aggregating their answers, try to eliminate unskilled participants. However, these methods do not work well in cases in which specific knowledge is required to solve the tasks, and where the distribution of participants is skewed, which can lead to unfair elimination of good participants and failure to exclude those who are unskilled.

III. Platform for Aggregated Data

In this section, we describe the outcomes of the Cityzen project in which we applied selective crowdsourcing. We begin by talking about the data model, then explain the architecture and details of the implementation, and finish by presenting the travel application.

A. Data Model

The Cityzen platform is intended to provide tourists with the ability to plan a trip that focuses on cultural heritage and to explore information about it in a spatial and temporal fashion. The information provided comes from distributed and heterogeneous sources, the aggregation of which poses various challenges cited and addressed previously in [6], where we created a data model for the Semantic Web [12] based on a Owl Ontology [13], and linked it to an external knowledge repository through Open Linked Data (LOD) [18]. The goals of having a Semantic Web-compliant data model are to allow other people and institutions to link their data to ours and for us to use their open data to expand the information available.

B. Architecture

The Cityzen platform is a web platform with data that are accessible freely through the Internet. To implement it, we modelled our knowledge base using Eclipse RDF4J [1], a powerful Java framework for processing RDF data. Its Workbench allowed us to create a repository that contained our data model and the instance data integrated from cultural heritage data sources. A second component of this framework, referred to as the RDF4J Server, allows online access to the repository through HTTP requests or SPARQL queries[2]. To make the data available, we installed the RDF4J framework in an Apache Tomcat Server [3]. Finally, we deployed everything on our server to make the data available to users. The data are accessible freely through SPARQL queries at the following link *http://datasemlab.ch:8080/rdf4j-server/repositories/CityZenDM* . Figure 1 shows the overall architecture of the platform.

C. Mobile Application

We created a mobile application to help tourists organize and engage in cultural heritage travel. The application consists of three modules:

- *"Cultural Heritage Viewer"* - allows users to find information about cultural heritage interests in a spatio-temporal way.

[1] http://rdf4j.org/
[2] https://www.w3.org/TR/sparql11-overview/
[3] https://tomcat.apache.org/

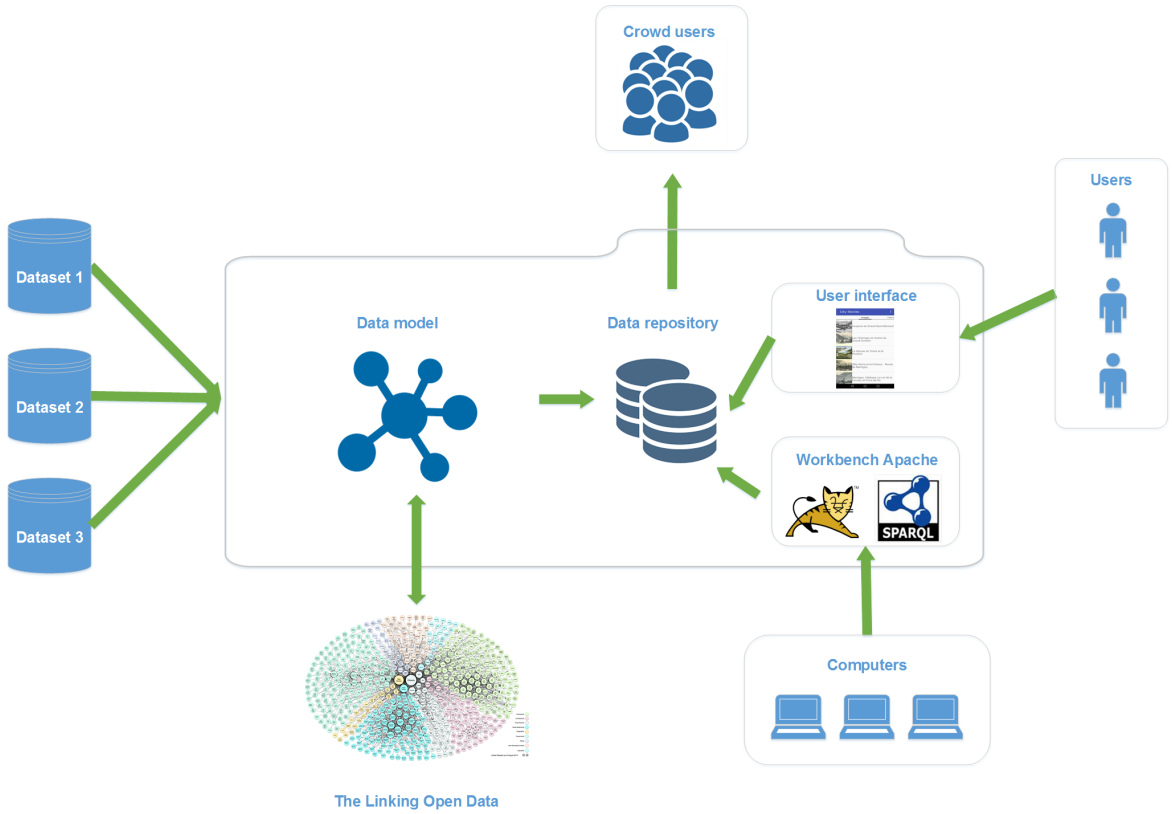

Fig. 1. Overall architecture of the platform

- *"Cultural Heritage Navigator"* - is connected with Google maps service and includes a notification feature that supports users during their sightseeing.
- *"Cultural Heritage Data Enhancer"* - includes the crowdsourcing feature to enrich and enhance the data.

Figure 2 shows screenshots of the mobile application.

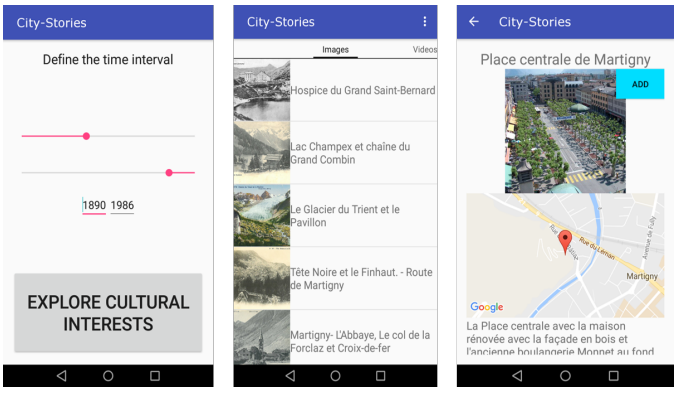

Fig. 2. Mobile Application

IV. Users Participation

Creating a data model that aggregates data about cultural heritage improves peoples ability to find interesting information and reduces the complexity of searches. However, integrating the data provided by different institutions includes some challenges and disadvantages.

The challenges are related primarily to data alignment, i.e., dealing with different data representations for the same pieces of information. Therefore, we developed a JAVA data converter that transforms each data representation into an RDF representation that complies with the data model.

However, the main issues with the data available are inconsistency and incompleteness. These are two issues that have different consequences in semantic web applications. By definition, missing information is tolerable, because the semantic web is based on the open-world assumption [10]. In contrast, inconsistency is an issue because different sources of data conflict, and thereby result in an inconsistent model that renders the data unusable. From an application perspective, the developers of applications can manage missing information based on our data model, while they cannot manage conflicting information, because it is unclear which information they should consider true. To address these two problems, we defined two approaches and developed them as separate features in our application.

A. Correcting Conflicting Information

Our data model is based on OWL, with which we can define restrictions for the schema relationships. When we integrate data from different providers, our platform can detect the presence of conflicts. This happens when there are more data

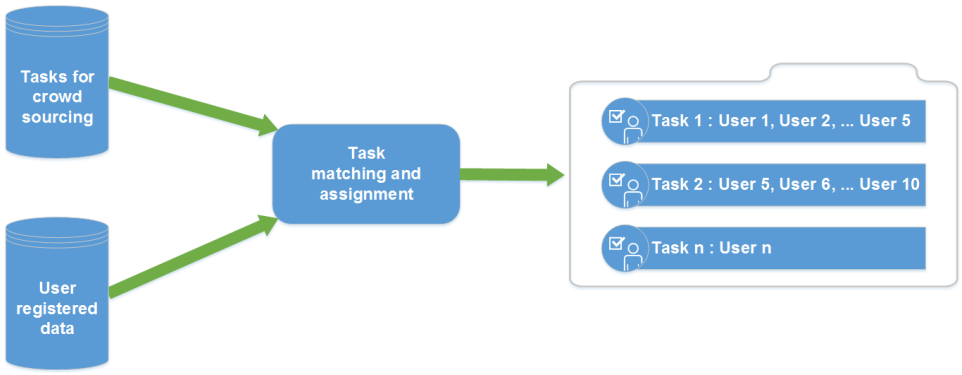

Fig. 3. Task matching and assignment process

associated through functional properties, which defines that only one instance of a datum can be present. We resolved this issue in the following way: data that create conflicts are not loaded in the data model, but are stored in a second database instead. These data are designed as small tasks for crowdsourcing that participants can be asked about later. Depending on participants preferences, such as location and cultural interests, tasks are directed accordingly to matched users to solve the issue. For example, a simple task could be a single choice question, such as, When did event A happen: in year B or in year C? Participants answers are stored and analyzed further before choosing the final answer that eventually resolves the data conflict.

B. Completing Information

As specified previously, missing information does not represent a conceptual error, but it can lead to limitations in the application's use. In our application, we designed micro tasks as short questions that asked participants to provide an answer to fill in the missing information in a data record. These small tasks are directed either to the user or appear while they browse a specific record that includes the issue of missing information. For instance, a participant can be asked to specify the date of an event, or its position via Google Maps.

We applied the gamification [11] concept, in which participants are awarded reputation points and badges depending on their level of contribution, to increase participants interactivity and their motivation to contribute. This approach is similar to the one used in contributions by Google Local Guides[4], but it does not have the disadvantage of distracting the user from its main purpose, as to provide an answers is optional.

The platforms policy for adding users contributions to the data model is to consider the data correct and insert them in the data model when at least 5 people have rated one of the options and the option rated most has been chosen 70% of the time.

After the first phase of testing the application, which was conducted by students in the tourism department of our university, we noticed that the approaches we adopted provided decent quality improvements. However, the percentage of conflicts solved and the information added were lower than we expected. Our hypothesis is that even though users are willing to participate, when they use the application, they prefer to focus primarily on its purpose, i.e., browsing and navigating the data and thus, do not always collaborate in the information amelioration process. This is due largely to their lack of motivation to contribute, even if gamification is used as incentive, which is a major issue in crowdsourcing in general. In the next section, we explain the approach that we took to improve this situation.

V. Solving Data Issues with Selective Crowdsourcing

Data in our repository usually lacked some aspect of the information, which might be a date, location, or incomplete title and description. Therefore, the crowdsourcing tasks were designed as a survey, following the style of online crowdsourcing platforms, such as Amazon Mechanical Turk (MTurk)[5] and SamaSource[6], which design micro-tasks for crowd participants referred to as Human Intelligence Tasks (HITs). Contrary to conventional crowdsourcing scenarios, in which crowd participants choose the questions, i.e., go to the platform and search for tasks to solve, in our case, we used the push-crowdsourcing technique [21], which models tasks by considering users information, and includes the tasks that are most relevant for each user.

In our case, we allowed users to specify basic profile information, such as age, location, place of origin, as well as more specific information, such as interest in topics such as geography, tourism, and history about the region in which they live. Having the users information and tasks ready for crowdsourcing, we analyzed and tried to match tasks with users, i.e., task modelling. This is important to obtain increased accuracy and user engagement, especially because of the specificity of the datasets that we used in this work, which require deeper knowledge on specific topics.

[4]https://maps.google.com/localguides

[5]https://www.mturk.com/

[6]https://www.samasource.org/

It is less challenging to design the HITs in the conflicting data scenario, as we do have a closed set of answers that derive from the integration conflict. As a result, the micro-task is a single choice question given to the users that asks them to select one of the options available.

Asking participants to solve micro-tasks is not the end of the process in crowdsourcing. Allowing every user to solve tasks can produce low quality results, as it happens to include lazy participants who do not consider the HITs seriously and therefore provide random answers. As a result, we applied quality control mechanisms to maintain the quality of the data derived from the crowdsourcing process. Some well-known quality checking strategies are aggregation techniques, which assign the same task to multiple users. We used the majority voting technique [19] by asking 5 users to perform the same task, in which the correct answer was the one with the most votes. Additionally, this is a good technique for pointing out lazy participants.

VI. Conclusions and Future Work

While the selective crowdsourcing experiment was running, we conducted a survey to ask crowd participants how they felt about the tasks they were assigned. The survey showed that the majority of the crowd was comfortable with the procedure, because the tasks assigned to them were related to their knowledge and interests. Moreover, in only a quarter of the total planned time for the solving conflicts test phase, the number of conflicts the crowd solved almost equalled the total number of existing conflicts. This confirmed our intuition and made us think that, in the future, researchers could develop techniques that allow people in real-time to contribute to data improvement. This would improve the current limitation of checking the contributions in offline mode, something that leads to delays on the updates.

References

[1] Literature and Artifacts. G. Thomas Tanselle Bennett Gilbert The Library Quarterly 2000 70:4, 504-506

[2] Arman Akhoondnejad, Tourist loyalty to a local cultural event: The case of Turkmen handicrafts festival, In Tourism Management, Volume 52, 2016, Pages 468-477, ISSN 0261-5177, https://doi.org/10.1016/j.tourman.2015.06.027. (http://www.sciencedirect.com/science/article/pii/S026151771500151X)

[3] Tourism statistics - characteristics of tourism trips. May 2017. URL: http://ec.europa.eu/eurostat/statistics-explained/index.php/Tourism_statistics_-_characteristics_of_tourism_trips#Further_Eurostat_information

[4] Prentice, R. *Tourism and heritage attractions.* 1993 pp.253pp. ref.168

[5] J. Ren, Y. Zhang, K. Zhang and X. Shen, "Exploiting mobile crowdsourcing for pervasive cloud services: challenges and solutions," in IEEE Communications Magazine, vol. 53, no. 3, pp. 98-105, March 2015. doi: 10.1109/MCOM.2015.7060488.

[6] Olivieri, Alex C. et al. Cityzen: a social platform for cultural heritage focused tourism. MEDES (2016).

[7] J. Bleiholder and F. Naumann. Conflict handling strategies in an integrated information system. In Proc. of WWW, 2006.

[8] McKnight, Patrick E., et al. Missing data: A gentle introduction. Guilford Press, 2007.

[9] Rahm, Erhard, and Hong Hai Do. "Data cleaning: Problems and current approaches." IEEE Data Eng. Bull. 23.4 (2000): 3-13.

[10] Drummond, Nick, and Rob Shearer. "The open world assumption." eSI Workshop: The Closed World of Databases meets the Open World of the Semantic Web. Vol. 15. 2006.

[11] Sebastian Deterding, Dan Dixon, Rilla Khaled, and Lennart Nacke. 2011. From game design elements to gamefulness: defining "gamification". In Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments (MindTrek '11). ACM, New York, NY, USA, 9-15. DOI: https://doi.org/10.1145/2181037.2181040

[12] Berners-Lee, Tim, James Hendler, and Ora Lassila. "The semantic web." Scientific american 284.5 (2001): 28-37.

[13] Antoniou, Grigoris, and Frank Van Harmelen. "Web ontology language: Owl." Handbook on ontologies. Springer Berlin Heidelberg, 2004. 67-92.

[14] Jiannan Wang, Tim Kraska, Michael J. Franklin, and Jianhua Feng. 2012. CrowdER: crowdsourcing entity resolution. Proc. VLDB Endow. 5, 11 (July 2012), 1483-1494. DOI=http://dx.doi.org/10.14778/2350229.2350263

[15] Xuan Liu, Meiyu Lu, Beng Chin Ooi, Yanyan Shen, Sai Wu, and Meihui Zhang. 2012. CDAS: a crowdsourcing data analytics system. Proc. VLDB Endow. 5, 10 (June 2012), 1040-1051. DOI=http://dx.doi.org/10.14778/2336664.2336676

[16] Tingxin Yan, Vikas Kumar, and Deepak Ganesan. 2010. CrowdSearch: exploiting crowds for accurate real-time image search on mobile phones. In Proceedings of the 8th international conference on Mobile systems, applications, and services (MobiSys '10). ACM, New York, NY, USA, 77-90. DOI=http://dx.doi.org/10.1145/1814433.1814443

[17] Johan Oomen and Lora Aroyo. 2011. Crowdsourcing in the cultural heritage domain: opportunities and challenges. In Proceedings of the 5th International Conference on Communities and Technologies (C&T '11). ACM, New York, NY, USA, 138-149. DOI=http://dx.doi.org/10.1145/2103354.2103373

[18] Bizer, Christian, Tom Heath, and Tim Berners-Lee. "Linked data-the story so far." Semantic services, interoperability and web applications: emerging concepts (2009): 205-227.

[19] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '08). Association for Computational Linguistics, Stroudsburg, PA, USA, 254-263.

[20] Pei-Yun Hsueh, Prem Melville, and Vikas Sindhwani. 2009. Data quality from crowdsourcing: a study of annotation selection criteria. In Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing (HLT '09). Association for Computational Linguistics, Stroudsburg, PA, USA, 27-35.

[21] Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudr-Mauroux. 2013. 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 (WWW '13). ACM, New York, NY, USA, 367-374. DOI: https://doi.org/10.1145/2488388.2488421

[22] Liliana Ardissono, Tsvi Kuflik, and Daniela Petrelli. 2012. Personalization in cultural heritage: the road travelled and the one ahead. User Modeling and User-Adapted Interaction 22, 1-2 (April 2012), 73-99. DOI=http://dx.doi.org/10.1007/s11257-011-9104-x

[23] Eide , Felicetti A, Ore CE, DAndrea A, Holmen J. Encoding cultural heritage information for the semantic web. procedures for data integration through cidoc-crm mapping. InOpen Digital Cultural Heritage Systems Conference 2008 Feb 25 (p. 47).

[24] Baca M. Practical issues in applying metadata schemas and controlled vocabularies to cultural heritage information. Cataloging and classification quarterly. 2003 Jun 1;36(3-4):47-55.

[25] Wu SC. Systems integration of heterogeneous cultural heritage information systems in museums: a case study of the National Palace Museum. International Journal on Digital Libraries. 2016 Nov 1;17(4):287-304.