Enriching the metadata of European colonial maps with crowdsourcing

Rouel de Romas

Vrije Universiteit, De Boelelaan 1105, 1081 HV Amsterdam, Netherlands
r.a.de.romas@student.vu.nl

Keywords: European colonial maps, crowdsourcing, digital humanities

Abstract. In this paper the effectiveness of crowdsourcing to enrich metadata about European colonial maps is tested. The repository of these European colonial maps contain small amounts of metadata about its sources. In the first part of this research, requirements for useful metadata about historical maps were identified by conducting an interview with an architectural historian. In the second part of this research includes participants who were asked to generate as many annotations about three European colonial maps, using an annotation tool called Accurator. Based on the requirements that were identified, the annotations of the participants were evaluated. The results indicate that the in most cases the annotations provided by the participants do meet the requirements provided by the architectural historian; thus, crowdsourcing is an effective method to enrich the metadata of European colonial maps.

1 Introduction

The study of Digital Humanities focuses on opportunities of using digital technology for humanities (Zhang et al., 2015). Often, activities that are being done in this study consists of creating ways to mine data out of historical sources (Ockeloen et al., 2013). A project, which specifically focuses on these kind of activities is ArchiMediaL[1].

ArchiMediaL is a project which specifically focuses on activities revolving around digital humanities, as well as cultural heritage. The goal of this project is to facilitate the automatic development and linking of metadata and image content, which can eventually be used for future research. What is pointed out in their project proposal[2] is that architectural sources will remain unexplored and unavailable for research if there are no methods for obtaining metadata about the architectural sources. One of the case studies that ArchiMediaL provides, regards the enrichment of the repository about European colonial architecture[3]. This

[1] https://archimedial.eu/

[2] ArchiMediaL: Developing Post-colonial Interpretations of Built Form through Heterogeneous Linked Digital Media. Unpublished(2016)

[3] http://colonialarchitecture.eu/

repository contains various sources of European colonial architecture, originated from the year 1850 to 1970, such as text documents, images and maps. With this problem and repository in mind, a viable data extraction method must be provided to enrich the metadata.

Crowdsourcing can be very useful for solving various types of problems with the help of human volunteers (Doan et al., 2011). With this method, large crowds can be reached in order for them to execute a specific task, which can also be done via the Web (Geiger et al., 2012). Withla (2009) suggests that crowdsourcing can be described as "a process of organizing labor, where firms parcel out work to some form of (normally online) community, offering payment for anyone within the 'crowd' who completes the task the firm has set". This method is beneficial due to the participants that can perform tasks at a low cost (Geiger et al., 2012). Withla elaborates on this advantage by stating that participants can be reached with various skills and expertise who are also able to execute tasks within a short time period. In some cases, the amount of skills, that the crowdsourcing participants contain, may be limited. Nevertheless, such crowdsourcing participants may still be willing execute tasks which are repetitive or require low amount of skill.

1.1 Scientific Contribution and Motivation

Opportunities have been created for the use of recently digitized historical sources due to the new digital media[4]; however, these historical sources contain small amounts of metadata. Thus, methods to extract metadata from these digitized historical sources must be created. For this paper, research will be done the historical maps that are included in the repository about European colonial architecture, since this historical source has not been analyzed yet. Furthermore, there are no significant research that has been done regarding the extraction of metadata about hisorical maps.

Historical maps can be useful for various reasons. Brovelli et al.(2012) argues that historical maps are a significant part of cultural heritage and can be considered as a valuable source of information for various purposes. Examples of purposes are historical and territorial research, architectural purposes, planning, archeology and demographic purposes. Marcucci (2000) goes more in depth into the importance of historical maps in the landscape planning process. According to Marcucci, to understand the current landscape, analysis must be done on the history of that landscape.

1.2 Research Question

In order to find out whether crowdsourcing will be effective for extracting metadata about European colonial maps, the following research question is formulated

[4] ArchiMediaL: Developing Post-colonial Interpretations of Built Form through Heterogeneous Linked Digital Media. Unpublished(2016)

for this research: "To what extent is crowdsourcing an effective method to enrich the metadata of European colonial maps?".

To answer the research questions, it is useful to know how to measure the effectiveness of crowdsourcing. In this research, effectiveness will be measured in terms of the usefulness of metadata. This makes it relevant to know what the requirements for useful metadata about European colonial maps, which leads to the sub question: "What are the requirements for useful metadata about European colonial maps?".

2 Related Work

Definition of metadata

Since there are various definitions for the term "metadata", in this research only one definition will be used as guidelines. According to Franks and Kunde (2006), the definition provided by The Minnesota Electronic Records Management Guidelines is described as one of the best definitions for the term 'metadata': "Metadata allows users to locate and evaluate data without each person having to discover it anew with every use. Its basic elements are a structured format and a controlled vocabulary, which together allow for a precise and comprehensible description of content, location, and value." The importance of metadata is explained by Alemu (2018). According to Alemu, metadata assist in locating cultural information by users. Moreover, due to large amounts of information on the web, its is preferable to have the cultural information systematically organized with the assistance of metadata.

Crowdsourcing and digital humanities

Various research has been done on crowdsourcing and digital humanities. In this section three, research papers will be discussed. In the first research paper, these opportunities and challenges are identified (Oomen et al., 2011). For this research, an analysis has been on multiple crowdsourcing projects. The opportunities regarding crowdsourcing have been classified in different crowdsourcing initiatives. Out of the six initiatives that are mentioned, only two initiatives are applicable for this research. In the first initiative called Classification, the users are asked to gather descriptive metadata related to object in a collection. In the second initiative called Co-curation, the inspiration or expertise of non-professional curators are used to produce information. Despite the opportunities, there are also challenges. One of the challenges is the quality of the information provided by crowdsourcing participants (Oomen et al., 2011). When working with an unknown network, it is difficult to obtain high quality information from the participants, since it is likely that non-experts are involved in crowdsourcing initiatives.

In the second research paper, it is argued that making collections of historical objects accessible online will create a connection to the past (Owens, 2013). The connection can be made by the way these objects are used, reused, explored, and understood. Crowdsourcing can contribute to this connection by enriching the data about these historical objects.

In a different research paper, two characteristics of crowdsourcing project with digital humanities are identified (Carletti et al., 2013). The first characteristic is the improvement of already existing collections with crowdsourcing. The second characteristic is the creation of new resources with crowdsourcing.

Other sources in the Colonial Architecture repository

Two types of sources in the Colonial Architecture repository have already been used for research. The first source is text documents containing architectural information about European colonies. Lo (2017) used Optical Character Recognition and Named Entity recognition to annotate relevant entities and terms that are used in the documents. Lo approached this task by building a pipeline in Python, which not only annotates relevant entities and terms but also links these entities and terms to knowledge that is available on the Web. The second type of source are images depicting buildings in European colonies. Brouwer (2017) tries to tackle the problem of interpreting architectural components that are depicted in the images through the use of crowdsourcing. The annotation tool Accurator (Dijkshoorn et al, 2013) has been used in order to gain metadata from the crowdsourcing participants of the research.

3 Research Method

The methodology of this crowdourcing research will be based on the methodology used for the Accurator tool (Dijkshoorn et al, 2017). The methodology contains four stages: orientation stage, implementation stage, execution stage and evaluation stage. In the orientation stage, the requirements concerning the useful metadata will be identified. Also, the objects used in the crowdourcing experiment and the participants are defined. In the implementation stage, the environment for the crowdsourcing experiment is set up. The tasks of the participants will be performed in the execution stage. In the evaluation stage, all the resulted annotations will be analyzed and evaluated.

3.1 Orientation stage

Requirements

One step in conducting this research is to gather the requirements for useful metadata about European colonial maps. The requirements are gathered from a architectural historian of the project ArchiMediaL, since it is the case study of ArchiMediaL this research is based upon. The architectural historian was informed about the objective of this research and asked the question: "How would you define 'useful' metadata about historical maps?". Knowing whether the participants meet the requirements formulated by the architectural historian of ArchiMediaL, will entail whether the information provided by the participants is usable or not. Furthermore, after obtaining the set of requirements, the subquestion of this research will be answered.

Participants

To be able to enrich the colonial architecture dataset of ArchiMediaL, I will implement a crowdsourcing approach. The objective is to involve 15 participants in the crowdsourcing experiment. The participants of this study will be people from different educational backgrounds in Amsterdam. The reason behind students with different educational backgrounds is that this will ensure different skills and expertise. Also, the difficulty of obtaining these participants is low.

3.2 Implementation

Tool

An environment will be set up for study in which crowdsourcing participants can provide annotations for the European colonial maps. This will be done with an annotation tool called Accurator. This tool has been created by Dijkshoorn et al.(2017), to obtain annotations about cultural heritage collections from crowdsourcing participants. Because Accurator initially was meant for cultural heritage collections, this tool will be very useful for acquiring annotations from European colonial maps.

Maps

The European colonial maps that will be used for this study is a map of the location Batavia, Ambarawa & Salatiga, and Sourabaya[5]. There are two reasons for choosing this location for this study. First, the locations are in comparison with other locations in the repository one of the most detailed maps in the repository, which means that more information can be extracted from these locations. Secondly, the depicted locations differ from each other in size. The largest location is the map "Ambarawa en Salatiga en Omstreken", since two cities and the surroundings are shown on the map. The map "Batavia Military Guide Map" is second in terms of size, since only the city Batavia is showm. The smallest location is the map "Plan of Sourabaya". This map is also a city; however, this city is smaller in size. All the three maps already contain a small list of metadata: title, location, scale, coordinates, publisher, date, location of the map, shelf location and rights[6][7][8].

[5] https ://figshare.com/articles/Maps_used_in_the_crowdsourcing_experiment/6725177

[6] http://colonialarchitecture.eu/obj?sq=id%3Auuid%3A25ca4ec8-7efc-4317-9426-842852b67112

[7] http://colonialarchitecture.eu/obj?sq=id%3Auuid%3Ac3c677a0-a30d-4118-be7e-9be0f9391c41

[8] http://colonialarchitecture.eu/obj?sq=id%3Auuid%3A88a2ffeb-c43d-4999-af93-354aa44e6fc3

Vocabulary

To help the participants of this study with annotating the European colonial maps, a structured vocabulary called Art & Architecture Thesaurus (AAT)[9] will be implemented in the annotation tool. The purpose of using the terms in AAT is to describe various objects in the field of art, architecture, decorative arts, material culture, archival materials. There are many facets in the vocabulary. The most relevant facet for this research will be the Objects facet. The reason being is that this facet contains terms that are related to tangible or visible things by created for the most part by humans. This is applicable for this study since the European colonial maps that is used depict tangible or visible things created by humans, such as buildings or harbors.

3.3 Execution stage

Pilot study

In order to identify design flaws and gain experience with participants and the tool, a pilot study, which is a small-scale version of a larger proposed study, will be conducted. According to Beebe (2007), also the feasibility of research methods and cost of research procedures can be examined when conducting a pilot study. The pilot study for this research will be done with only one or two participants since the pilot study is a smaller scale study.

Tasks

The tasks that the participants will perform are straightforward. Each participant will be shown maps mentioned in the Implementation stage. The goal for the participants is to annotate as many information as possible. There are no restrictions given to the participants for giving annotations, since these will eventually by evaluated afterwards. The duration of the tasks will not have restrictions as well to ensure a large amount of annotations.

Guide on using the annotation tool

Since the participants of this study may not be specialized in the field of history or cartography, it is important to let the these participants go through their tasks successfully. Thus, a guide [10] will be given to the participants. In this guide, the participant will go through the annotation tool step by step. The guide consists of four steps. The first step contains instruction on how to register in the Accurator, since annotation can only be made when having an account on the Accurator server.

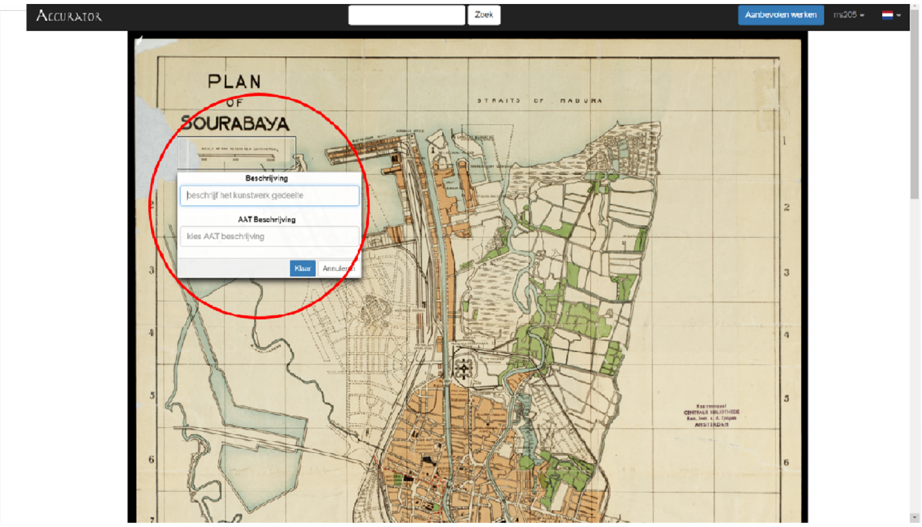
[9] http://www.getty.edu/research/tools/vocabularies/aat/about.html#history

[10] https://figshare.com/articles/Guide_on_how_to_use_the_annotation_tool_Accurator/6716249

The second step contains instructions on how to choose the map that will be used for this study. Choosing the map is not straightforward due to the inclusion of other images, not relating to this research, in Accurator.

The third step contains instruction on how to create annotations. There are two ways of providing annotations on an image. The first approach is to select a specific spot on the image by creating a box around this specific spot on the image (Figure 1). After selecting a specific spot, the participant can give a description about the selected spot with an AAT description or a non-AAT description. The second approach is to provide descriptions of the image that cannot be observed directly from the maps (Figure 2).

Fig. 1: First annotation function

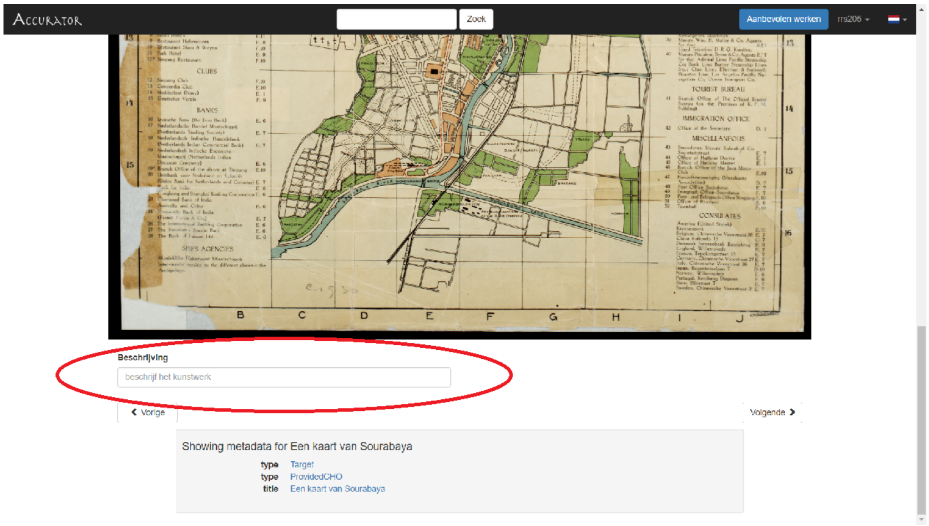


The fourth step contains a brief questionnaire on the experience of the participant with the annotation tool and the guide. The goal of this questionnaire is to collect the experiences of the participants and improve the guide on how to use the annotation tool. The participants have to provide up to three positive features and up to three negative features of the annotation tool. Furthermore, the participants must give their opinion on the difficulty of the task. Also, if willingly, the participant can give advice on how to improve the guide in order for future participants to go through the guide more easily.

3.4 Evaluation stage

After the experiment is conducted, an evaluation will be made on the generated annotations of the crowdsourcing participants. This evaluation consists, firstly, of manually counting the total amount of annotations. This part of the analysis

Fig. 2: Second annotation function



Beschrijving

beschrijf het kunstwerk

Showing metadata for Een kaart van Sourabaya

type	Target
type	ProvidedCHO
title	Een kaart van Sourabaya

will show the characteristics of the resulted annotations. The second part of the evaluation is analyzing whether the annotation of the participants meet the requirements that resulted from the interview. Each annotation will be matched to one of the categories of useful metadata about European colonial maps. When at least 70% of the requirements are met in this study, a crowdsourcing approach for enriching the metadata of European colonial maps can be considered effective.

4 Results

In this section all the results will be described. All the files relating to the results are published on Figshare[11]

4.1 Requirements for useful metadata

As described in the research method, the requirements for useful meta data will be identified in order to identify whether the crowdsroucing participants meet the requirements of useful metadata. To identify the requirements, contact has been initialized with a historian who is active in the project ArchiMediaL[12]. By profession the person is an architectural historian and is currently contributing in the Architectural History department of ArchiMediaL. The request given to the architectural historian is to answer to question ”How would you define ’useful’

[11] https://figshare.com/projects/Enriching_the_metadata_of_European_colonial_maps_with_crowdsourcing/35690

[12] https://figshare.com/articles/Questions_for_gathering_requirements_for_useful_metadata_about_colonial_maps /6725123

metadata about historical maps?" In response to this question, the architectural historian provided a list of useful metadata. After observing this list of 12 types of metadata, a categorization could be made out of it, which is listed below in Table 1. There are five categories identified. The first category is background information. These types of information concern, for example, date and author. The second category contain metadata that corresponds to the scale and type of the map. The third category is Content, which entails metadata about the content of the map. The fourth, the category Perspective involves metadata about the multiple perspectives the map has. The last category is Specific, which contain metadata such as depictions within or aside the map.

Table 1: Requirements for useful metadata

Background Information	**Map**	**Content**	**Perspective**	**Specific**
- Date - Author - Author's nationality/ origin - Commisioner - Map's content	- Scale - Type of map (e.g. political, geographical, etc.)	- Text in the map (all labeled elements) - Elements used (e.g. streets, forest, waterways, buildings, types of buildings, fortifications, etc.)	- Abstraction level (relatively realistic, distorted,) - Perspective (top view, perspective, 3D, etc.)	- Specific Elements (coat of arms, skyline view, depictions within/aside the map)

4.2 Pilot Study Results

The pilot study is done by two participants[13]. One participant is a male person born in 1994 and the other participant is a female person born in 1999. They were requested to annotate only one map out of three maps that were provided. The male participant provided 8 annotations while the female participant provided 16 annotations. The two participants gave feedback on the annotation tool and the guide that is given to them. Even though, the male participant found it easy and straightforward how to add annotations, he stated that he had difficulties coming up with metadata. Moreover, he believed that the instructions given to him needed to be more easy to follow and terms that not everybody was familiar with need to be avoided. On the other hand, the female participant did not have difficulties providing metadata. Her only concern was the overlapping annotations when there are one or more depictions on the map are close to each other when. Only the feedback of the male participant was applied, as only the guide of the crowdsourcing experiment could be altered.

[13] https://figshare.com/articles/Survey_Results_of_the_Pilot_Study/6726152

4.3 Participants

In total 15 participants contributed to the research. All of the participants were students ranging from the birth years 1993 to 2000. With its median being the birth year 1997. The distribution of male or female is not equal, since there are 10 male participants and 5 female participants. The educational background of the participants are diverse: Information science; Economics; Communication and Multimedia Design; English Literature; Science Business and Innovation; Civil Engineering; and Application Developing.

4.4 Annotations

In total, 458 annotations were provided by the participants. A participant provided, on average, approximately 30,5 annotations. All the annotations are published on Figshare[14]. Each participant was requested to annotate three maps. The distribution of all the annotations among the three maps used for this research can be seen in Table 2. The first map was the location Batavia. In total, 167 annotations were made on this map. The second map depicted the locations Ambarawa and Salatiga. In total, 135 annotations were provided for this map. The third map depicted the location Sourabaya. In total, 156 annotation were made by the participants for this map. Using the table of useful metadata, the distribution of all the annotations among the categories can be seen in Table 3[15]. Most of the annotations concerned metadata about the content of the map, which are 208 (45,2%) annotations. The least amount of annotations concerned the metadata about specific, which are 19 (4,1%) annotations. The auto correction function with the AAT, described in the research method, is barely used; only 2 of the 440 annotations were produced with the autocorrection tool.

Table 2: Annotations per map

	Batavia	**Ambarawa & Salatiga**	**Sourabaya**	**Total**
Amount Annotations	167	135	156	458
Average annotations participant	11.1	9	10.4	30.5

What was also noticeable about the annotations is that not all the annotations fit into a certain category; as a result, another category is made called ”Other”, which contains 88 annotations[16]. Furthermore, not all the annotations of the participants were correct; in total there were 24 annotations that can be

[14] https://figshare.com/articles/Annotation_in_RDF_language/6725084

[15] https://figshare.com/articles/Annotation_distribution_of_the_three_maps/6726176

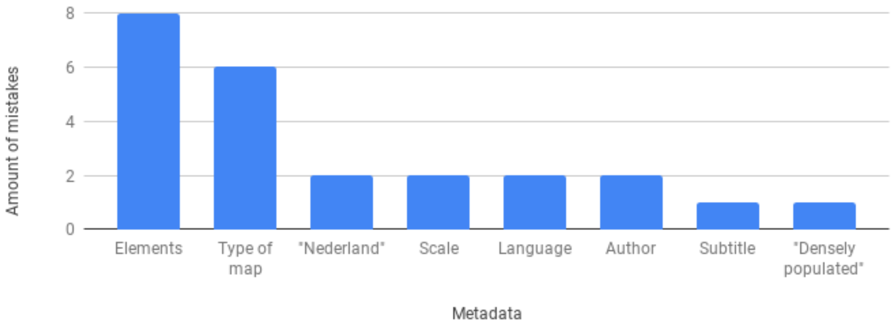
[16] https://figshare.com/articles/Metadata_in_the_category_Other_/6726038

Table 3: Annotation distribution of all maps

	Background Information	**Map**	**Content**	**Perspective**	**Specific**	**Other**	**Mistakes**
Batavia	21	20	72	9	4	35	6
Ambarawa & Salatiga	13	21	49	13	8	19	10
Sourabaya	4	9	87	8	7	33	8
Amount of annotations	38	50	208	30	19	87	24
Percentage of the total amount	8.3%	10.9%	45.2%	6.6%	4.1%	19.2%	5.2%

considered false (Figure 3)[17]. The most common mistakes were about metadata about the elememts used in the maps and the type of maps.

Fig. 3: ”How difficult was it to come up with annotations?



4.5 Feedback

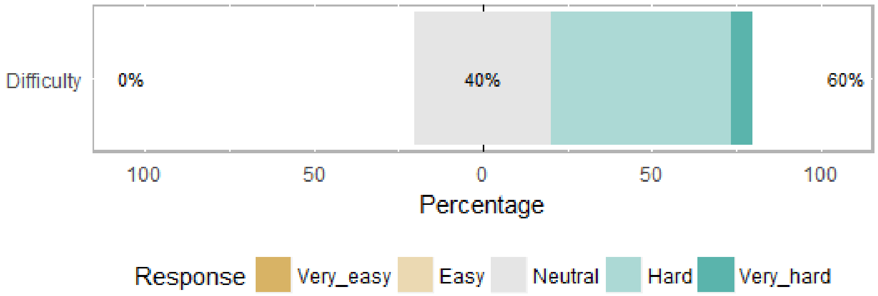
The participant of the study also provided feedback on their experience with annotating the three maps. Their feedback consisted of four categories: positive feedback, negative feedback, difficulty with annotations, and advice. In general, the crowdsourcing participants found the application relatively easy to use. Regarding the negative feedback, participants in general stated that they had difficulties observing smaller objects on the map, since the map is showed on a relatively large scale. Also when an annotation is entered by the participant, there is no possibility to edit this annotation. As advice, some participants mentioned adding a function where you could zoom in to the images, since the maps

[17] https://figshare.com/articles/Annotations_categorized_as_false/6726188

are shown on a relatively large scale. Furthermore, some participants gave as advice to make the guide on how to use the annotation more appealing.

The crowdsourcing participants also provided their opinion on the difficulty of coming up with annotations (Figure 4). 60% of the participants had difficulties coming up with annotations about the European colonial maps. 53.3% of the participants stated that it was ”hard” to come up with annotations and 6.7% of the participants stated that it was ”very hard” to come up with annotations. On the other hand, 33.3% neither found it hard or easy to come up with annotations about the European colonial maps. None of the participants found it easy to come up with annotations about the European colonial maps.

Fig. 4: ”How difficult was it to come up with annotations?



5 Discussion

The resulted distribution of the annotations shows that in all three of the maps the most annotations relate to metadata about the content of the map. The reason behind this is that the crowdsourcing participants found it relatively easy to annotate the content of the map since these kinds of metadata are the most easy to observe due to the large coverage it has on all the maps. The least amount of annotation relate to metadata about specific elements on the map. The reason for this may be the small amount of specific elements that is depicted on the map. Most of the participant that annotated specific elements were annotations about stamps and writings on the maps, which is not often the case for maps. When looking at all the annotations, most of the requirements for useful metadata have been met, which is 9 out of the 12 requirements (75%); only a couple of metadata have not been annotated. These metadata are: author, author's nationality/origin, commissioner. These type of metadata have not been annotated due to the absence of these metadata on the three maps.

Most of the participants found it difficult to come up with annotations. One of the reasons most of the participants found it difficult to come up with annotations, was due to their belief that they lack the knowledge to provide useful annotations. The lack of knowledge can be seen in the false annotations that were made. This problem relates to the difficulty of assuring that information is of high quality described by Oomen and Aroyo (2011). On the other hand, from the 40% that was neutral in giving their opinion on whether it was difficult or not, provided in some cases more detailed annotations in terms of description of the annotation.

What is observed from the results gained from this research is that most of the requirements identified by the architectural historian are met. However, it is difficult to determine how detailed most of the annotations were, even though there is an autocorrection function that can determine how detailed an annotation is. A large amount of the annotations were provided without the help of the autocorrection function available in Accurator. An important factor for the low amount of annotations provided with the AAT autocorrection function may be the unawareness of the accessibility of this function by the crowdsourcing participants. As seen in Figure 2, the AAT autocorrection function is placed below the function that allows the participants to write their own annotations without the assitance of the AAT autocorrection.

Another limitation of this research is that it is not safe to state that the requirements identified by the architectural historian are the only requirements that determine whether data is useful metadata about European colonial maps. Another source of uncertainty is the evaluation method of this research. All the requirements were matched manually by one person, which may have affected the results for this research.

Further research could be done on acquiring more requirements for useful metadata about European colonial maps from various sources; for example, literature that has not yet been identified for this research or experts in other fields such as cartography or experts specialized in the history of post-colonialism. Also, a further study with more focus on the involvement of participants specialized in the field of European colonial maps is suggested to give correct and more insightful metadata about the European colonial maps.

6 Conclusions

The aim of this research was to determine whether crowdsourcing is an effective way to enrich the metadata of European colonial maps. The results of this research show that crowdsourcing can be an effective way to enrich the metadata of European colonial maps. 75% of the requirements that determine whether data is useful or not, has been met by the annotations provided by the crowdsourcing participants. Only metadata about the author, author's origin and commissioner were not identified. The findings in this research will be of interests for people who want to gain information about metadata of historical maps or digital humanities projects like ArchiMediaL, who may want to involve participants who do not possess the skills or expertise in a certain field.

References

1. Doan, A., Ramakrishnan, R., & Halevy, A. Y. (2011). Crowdsourcing systems on the world-wide web. Communications of the ACM, 54(4), 86-96.
2. Geiger, D., Rosemann, M., Fielt, E., & Schader, M. (2012). Crowdsourcing Information Systems-Definition Typology, and Design.
3. Dijkshoorn, C., De Boer, V., Aroyo, L., & Schreiber, G. (2017). Accurator: Nichesourcing for Cultural Heritage. arXiv preprint arXiv:1709.09249.
4. Oomen, J., & Aroyo, L. (2011, June). Crowdsourcing in the cultural heritage domain: opportunities and challenges. In Proceedings of the 5th International Conference on Communities and Technologies (pp. 138-149). ACM.
5. Ockeloen, N., Fokkens, A., Ter Braake, S., Vossen, P., De Boer, V., Schreiber, G., & Legne, S. (2013, October). BiographyNet: Managing Provenance at Multiple Levels and from Different Perspectives. In LISC@ ISWC (pp. 59-71)
6. Zhang, Y., Liu, S., & Mathews, E. (2015). Convergence of digital humanities and digital libraries. Library management, 36(4/5), 362-377.
7. Beebe, L. (2007). What can we learn from pilot studies? Perspectives in Psychiatric Care, 43(4), 213-218.
8. Carletti, L., Giannachi, G., Price, D., McAuley, D., & Benford, S. (2013). Digital humanities and crowdsourcing: An exploration. Museums and the Web.
9. Brovelli, M. A., & Minghini, M. (2012). Georeferencing old maps: a polynomial-based approach for Como historical cadastres. e-Perimetron, 7(3), 97-110.
10. Brouwer P. (2017). Crowdsourcing viability for Archimedial. Unpublished manuscript
11. Lo, G. (2017). Creating a Colonial Architecture Pipeline. Unpublished manuscript
12. Owens, T. (2013). Digital cultural heritage and the crowd. Curator: The Museum Journal, 56(1), 121-130.
13. Franks, P., & Kunde, N. (2006). Why metadata matters. Information Management,40(5), 55.