Personalization in Crowd-driven Annotation for Cultural Heritage Collections

Chris Dijkshoorn[1], Jasper Oosterman[2], Lora Aroyo[1], Geert-Jan Houben[2]

[1] The Network Institute, Dep. of Computer Science, VU University, Amsterdam, The Netherlands
{c.r.dijkshoorn, l.m.aroyo}@vu.nl
[2] Web Information Systems, Delft University of Technology, The Netherlands
{j.e.g.oosterman, g.j.p.m.houben}@tudelft.nl

Abstract. Many cultural heritage institutions are confronted with a big challenge when it comes to adapting the process of registration, annotation and digitization of their collections to meet the new technological demands for providing their collections online with Web and mobile technologies. With limited funding and limited professional resources, annotation is often lagging behind significantly in terms of quantity and level of detail. Quantity as well as quality of the annotations is the key to finding relevant objects by various groups of users in online collections. As many experiments with cultural heritage data online show, human computation can be useful for collecting large quantities of data. Various of these experimented with a crowd of lay people to help with the annotation of cultural heritage artefacts. However, a common concern relates to the resulting quality. The current trend appears to focus on supporting 'the crowd' to achieve the desired level of quality. In this paper, we discuss the results of the first phase of an evolutionary system design performed in collaboration with the Rijksmuseum Amsterdam, as part of the SEALINCMedia project. By providing personalized support to the individuals and groups in the 'crowd' when annotating museum collection objects, we also support the integration of the annotation result in the RMA content management system. The contribution of this paper is twofold: (1) requirements specification for a crowdsourcing annotation process and (2) design of the personalization functionality in a crowdsourcing annotation system.

Keywords: Personalization, User Modeling, User Profiling, Cultural Heritage, Crowd-based Annotation

1 Introduction

Cultural Heritage (CH) institutions collect and preserve important societal artefacts and have the task to make these artefacts available to the general public. For this they generally prepare expositions and provide on-site facilities where the public can explore and learn about the physical artifacts in well-curated context. Following the general trend to go digital, most of the CH institutions are now also providing online

adfa, p. 1, 2011.

access to their collections [1]. Annotations made by professional curators of collection objects typically are done to meet the needs of other art-history professionals, rather than the ones of the general audience. Moreover, these annotations are typically covering a basic set of descriptive aspects, which do not allow for generating automatically interesting links between objects, so that each collection object can be presented in the context of other related ones (similar to the presentation in an exhibition) [2]. This means that the current annotations (i.e. textual descriptions of aspects or the entire object) need a significant adaptation to serve the needs of the general public.

Traditionally, the role of the CH institutions is to provide a quality stamp for the information provided by them, including the annotations of their collection objects. Thus, one of the main responsibilities of the museum cataloguers is to provide correct and suitable annotations to describe accurately each object in their collection. However, this process is usually time consuming and requires a large amount of highly-qualified human effort. Many CH institutions have turned to the 'crowd 'to face the quantity challenge [3]. Now that many already have gained a significant experience with using crowdsourcing methods to collect large quantities of data, the next challenge is how to ensure an acceptable quality of this result.

The Rijksmuseum Amsterdam (RMA)[1] has a collection of about 600.000 prints, which makes it one of the largest collections of prints in the world. Up till 2007 the only way to view some of those prints was to visit the RMA study room or wait for them to be included as a part of an exposition. Prints in expositions, however, could only be presented for a limited period of time due to the fragile nature of paper. To improve the access to their prints the Rijksmuseum started in 2007 the project **Print Room Online**. Its main goal is to realize the registration, annotation and digitization of the print collection. For nearly 5 years the team of 6 cataloguers, a photographer, a curator and a project manager registered and digitized about 146.000 prints. Even though the process has been optimized over the years, the capacity for registration and digitization is simply not large enough to finish the prints within a reasonable amount of time. Moreover, the RMA catalogers not always have the sufficient expertise in all the domains covered by the prints in this collection. In this context, in collaboration with the Rijksmuseum Amsterdam, and as part of the SEALINCMedia project[2], we investigate *how personalization can be applied in a crowdsourcing process for (1) reaching the optimal efficiency of each individual in the crowd, (2) maintaining a highly motivated community as a whole, and (3) achieving the right level of quality of the crowdsourcing results*.

We take an explorative and evolutionary approach: we have designed a system that allows configuration of various annotation support strategies and to measure and diagnose their performance. First, we analyze the current annotation process (Section 2) and derive a new process design that includes external annotators (Section 3). Further, we propose a design for a crowdsourcing annotation system by first considering strategies and methods for crowdsourcing annotation and its personalization functionalities (Section 4) and then presenting details of the design (Section 5). In Section 6 we

[1] http://rijksmuseum.nl

[2] http://www.commit-nl.nl/projects/socially-enriched-access-to-linked-cultural-media

discuss the evolution of the design. We conclude with a discussion relevant for similar CH institutions. The ultimate goal is to help RMA to understand and further exploit the role of external experts in the RMA annotation process.

2 Current Print Room Annotation Process

The overall challenge of this work is to find the optimal combination of the results produced by a crowd of knowledgeable amateurs and the efforts of museum professionals for increasing the annotation capacity in museums. We analyzed in detail the RMA Print Room Online project and identified the steps in the workflow suitable to delegate to the crowd.

As we noted earlier, the Rijksmuseum Amsterdam has a huge annotation backlog of prints and an annotation process that is solely aimed at supporting art-historians and other professionals inside and outside of the museum to find collection objects. The current workflow of the Print Room Online project (Figure 1) is highly standardized and well documented, and it assigns main annotation tasks to professional catalogers which is different from the general RMA workflow where the curators are solely responsible for the entire workflow except the digitization. It consists of four consecutive steps as depicted in Figure 1, namely register objects, provide basic annotation, digitize annotated objects, and finally perform quality control of the complete result. The workflow starts with selecting a box of prints. Each box contains all prints by the same creator. All the boxes are organized in an alphabetical order and selected for registration, annotation and digitization in this order. In the following four subsections we describe in detail the activities performed on each of those steps:

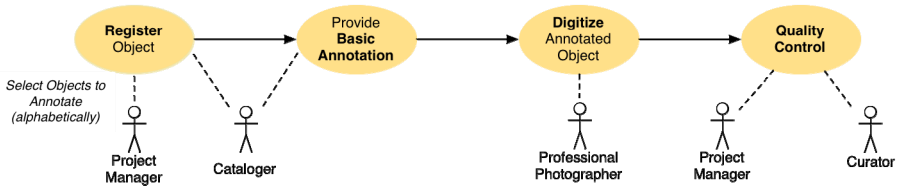

Fig. 1. Print Room Online project workflow

2.1 Register Object

The catalogers register each object from the box in the collection management system Adlib by entering its information (e.g. acquisition date, creator's name, name of the object, price and previous owner) from the RMA inventory books. Those inventories contain entries for each object at the time of its acquisition.

2.2 Provide Annotation

Next is providing the actual annotation for the registered object. As the catalogers work on one box of prints at a time, meaning prints created by the same artist, rele-

vant literature regarding the particular artist is therefore retrieved before the actual annotation process begins. These are the main sources used as a basis for the annotations. In case an individual print requires more information than what is available in those sources, more relevant books are retrieved. All the books (sources) used in the annotation process are recorded in Adlib as reference. In this step the cataloger spends about 15-20 minutes to annotate the most important basic aspects of the object by filling in fields in selected Adlib views, e.g. physical dimensions, date of creation, provenance information and subject matter. Subject matter typically concerns the depicted event, people, place and additional objects. Generally, catalogers are able to add general or specific descriptions to all of the annotation fields. However, it does happen that aspects of the subject matter or particular domains are outside of their expertise, and then external experts are consulted when available.

2.3 Digitize Object

Upon completion of the annotation of the entire box it is handed over to a professional photographer for digitization. Roughly about 150 objects a day are digitized in different resolution sizes.

2.4 Quality Control

Correctness and high quality of the information describing collection objects is essential for the RMA. The Print Room Online project leader does the majority of the quality control, e.g. checking that the right fields are filled according to the RMA documentation, as well as the correctness and consistency of the description fields and the bibliography. The project leader spends roughly 8 hours per week on quality control, which results in about 150 objects per week. The curator is responsible for checking the *title* and *description* fields. Problems and changes are further communicated to the catalogers to ensure optimization of the process.

3 Including the Crowd in the Annotation Process

The analysis of the four steps in the current annotation process has been done through interviews with RMA professionals and studying the annotation guidelines and Adlib software. Based on this, requirements have been elicited in order to identify additional steps in the workflow that can be performed by the general audience and/or by a crowd of interested and knowledgeable amateurs. In this section we outline the steps of an adapted annotation workflow that involves users external to the RMA and following in Section 4 we present the design of the system to support them.

The envisioned update to the workflow is based on the analysis of the current annotation process and intensive discussions with the RMA staff about the potential use of the crowd. We anticipate that crowd users require different support mechanisms compared to the professionals of the RMA and we derive opportunities and conditions that can be addressed for successfully involving external annotators in the process.

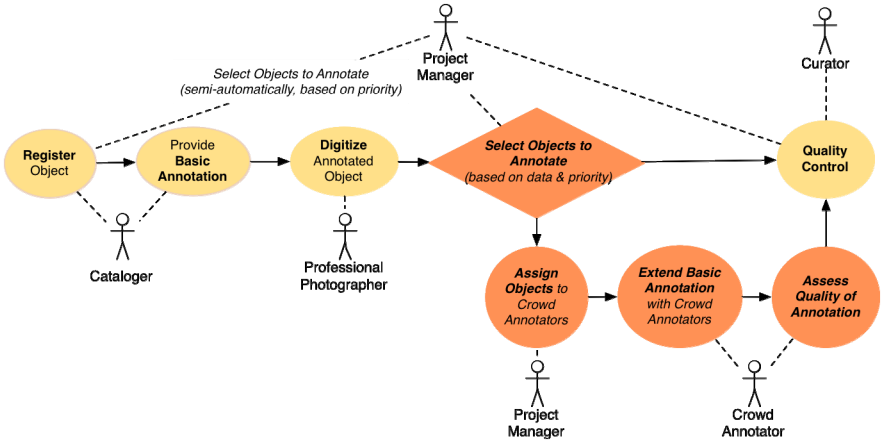

Fig. 2. Adapted Print Room Online workflow with crowdsourcing

The main additional steps in the workflow, illustrated in Figure 2 and addressed in more detail in Section 4, are the selection of objects that are given to the crowd to annotate, and then for each of those objects three separate steps: the assignment of the object to crowd annotators, the actual crowd annotation, and the assessment of the quality of the provided annotation.

4 Strategies and Methods for a New Annotation Support

In this section, we take a more detailed look at the adapted workflow. We provide an overview of strategies and methods that can be involved in the steps that concern the crowd and that are ultimately aimed at an effective interplay between RMA employees and external annotators from the crowd for the benefit of a more effective annotation process for the RMA. The next sections describe in more depth the four steps in orange of Figure 2. Section 4.1 describes the *Select Objects* step, 4.2 the *Assign Objects* step, 4.3 the *Extend Basic Annotations* step and 4.4 the *Assess Quality of Annotation* step.

4.1 Select Objects for Crowd Annotation

In order to maximize the effect of the crowd annotation for the annotation coverage of the whole collection, different strategies can be followed to select which objects to hand to the crowd for further annotation. One of the considerations in selecting objects for crowd annotation is to identify sub-collections with a significant amount of objects in need of annotation. For example, this could be related to priorities (needs) in terms of special exhibitions that are planned or the desire to further develop certain topic areas: which parts of the collections will be explored first? A second considera-

tion for object selection strategies could look at objects that have a low coverage of annotations or have annotations which can benefit from further refinement. This relates to the opportunity and feasibility of giving the object to the crowd for annotation: is there sufficient information available to expect a possible crowd annotation? Third, strategies could consider whether sufficient general information about the object is available to create the possibility to later match the item to an annotator with the desired expertise. Since in the crowd not all are equal, for the quality goals it is relevant to know whether it is feasible to give the object to relevant crowd annotators: is there sufficient information available to expect a possible relevant annotator? This first new step in the workflow is executed based on decisions and strategies like these in order to obtain an effective contribution of the crowd.

4.2 Assign Objects to Crowd Annotators

External crowd annotators, who are for example invited to participate in an online annotation environment, could have specific areas of expertise, and therefore they might benefit from so called 'task routing'. Intelligent task routing is defined as "matching people with appropriate tasks" and has shown to increase up to roughly four times the edits made by volunteers on Wikipedia articles, compared to a scheme that presents them with random articles [4]. With the aid of user profiles recording an external annotator's level of expertise, strategies could recommend appropriate prints to annotators, thereby increasing the productivity of the contributors. For example, a botanist will have more to contribute to an object with a depiction of a flower, then to an object depicting a naval battle. This step in the workflow is aimed at the identification of relevant annotators for an object, and strategies in this step could consider what is known about the object, about the already available annotation, about the annotators, and the match between them.

4.3 Extend Basic Annotations with Crowd Annotations

Opening up the annotation process will mean that external people with different areas of expertise can add annotations to RMA collection objects. Depending on their level of expertise it is likely that they would able to add detailed descriptions to specific annotation fields. This gives a number of opportunities to maximize the overall performance. First of all, a *selection can be made which additional fields* (besides the basic annotation fields filled by the catalogers) external experts could annotate. In this respect, to exploit the best of each expert it can be wise to match appropriate fields to a particular expert: for example, the annotation field regarding persons is prioritized for an annotator who is an expert in recognizing historical characters.

The use of professional jargon in the annotation process can hinder the participation of non-specialist users. An example is iconography, for which the RMA uses a specialized vocabulary called Iconclass[3]. Iconclass uses codes to indicate which concepts are depicted on a print. These codes can denote diverse concepts, ranging from

[3] http://iconclass.org

'*windmills*' (iconclass code: 47D31) to the '*Virgin Mary*' (iconclass code: 71C1(+2)). Here, we identify another opportunity to support the external annotators by giving them *recommendations concerning appropriate annotations*. Several strategies are considered here. It can be done using auto-completion, based on the vocabularies available at the RMA, i.e. matching a list of terms from the VIAF[4] or RMA People vocabularies. This way ensures that external annotators use concepts from existing vocabularies (when possible). The scope of the entities recommended by the auto-complete function could be further narrowed down by taking the previously filled in annotation fields into account. Given an artist name, other people that collaborated with this artist or work in the same period could be recommended (in addition to locations, temporal information and related events). These recommendations can be refined whenever the user adds more information to the record.

Additionally, the recommendation of concepts can combined with supporting annotators by *considering their preferences and characteristics*, thus using strategies to present supporting documentation and information in different ways to them. For example, professional catalogers have different sorting preferences for annotation fields [5]. It would be interesting to see if external annotators share these preferences by exploring the effect of a number of sorting methods. Choosing the correct concept from a recommended set can be aided by showing information that allows disambiguating of concepts and also selecting different levels of detail corresponding to the level of expertise of the user.

Finding the right sources and documents to provide supporting content for the annotations could be critical for the quality and the correctness of the annotations. Thus, (*personalized) search* can improve the effectiveness of both professionals and external experts. Moreover, using semantics in the search strategies can help finding also related documents and exploring the domain for users not familiar with all the sources (both 'professional' as well as open web resources). Different strategies could be envisioned that help provide the annotators in reducing their possible knowledge gap, for example by looking at the creator, the subject or similar paintings already annotated. The recommended sources can either be available online, or point to books or paragraphs of books. Similar prints can be shown, to provide annotators with examples.

4.4 Assess Quality of Annotations

The RMA has extensive quality control process to ensure the correctness of the annotations. Not all annotations are inspected because it is time consuming. This global quality control process will remain in place even when external annotators will be involved: it could be expected that over time certain quality control policies regarding the annotation extensions performed by externals could be adjusted, if the inspection of the assessed quality provides evidence for this, and the RMA develops a certain trust in the crowd process. Therefore, in order to gather evidence of the external annotation performance, the crowd annotation process includes a quality assessment step.

[4] http://viaf.org

This step is aimed at collecting data about the perceived quality of external annotations, for example through peer review, majority voting, or occasional professional verification. Thus it becomes possible to make an assessment for each annotator whether it is appropriate to place trust in him or her, as well as to make assessments of the feasibility of crowd annotation for certain topics or parts of the collection. The data thus collected can be used for adjusting and refining the previously applied strategies in the crowd annotation process itself and to help in the global quality control process of the whole Print Room Online Process.

5 Design of Personalized Annotation System

In this section we discuss the data (Section 5.1) used by the components (Section 5.2) in the architecture of the annotation system supporting strategies from Section 4. We focus on the component interaction and implementation details of the components specifically related to personalization.

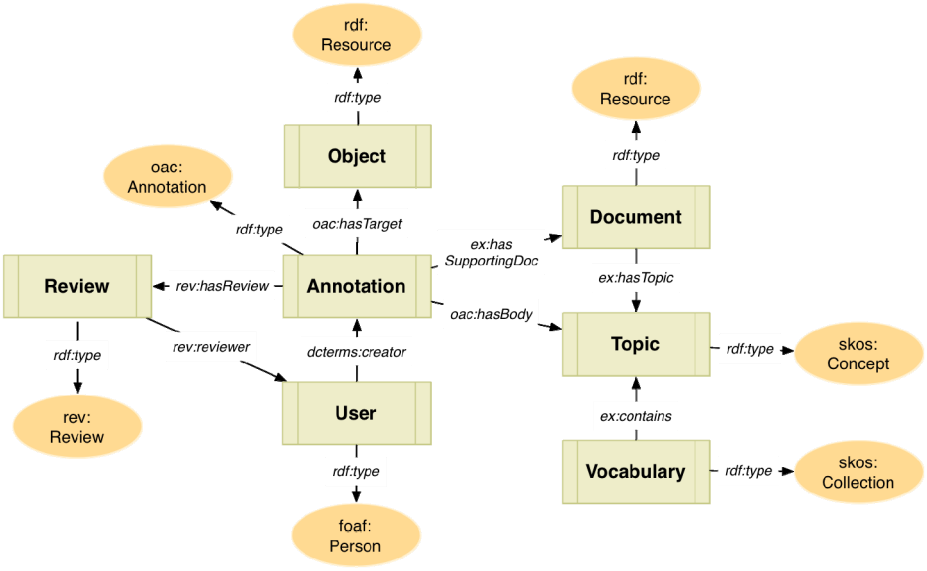

Fig. 3. Entities and their RDF representation

5.1 Entities

The system consists of seven conceptual entities which are shown in Figure 3. We use these conceptual entities to explain the system. In the implementation of the system the data in each entity is stored in RDF. Figure 3 also shows with what RDF type each entity is represented and how the relations between these entities are stored in RDF.

We reuse a number of vocabularies: Dublin Core terms[5] (dcterms) for metadata, Friend-of-a-Friend[6] (foaf) for user information, Open Annotation[7] (oac) for annotations, Review[8] (ev) for reviews and SKOS[9] (skos) for topics and vocabularies. The prefix ex indicates our own namespace for relations that are not present in existing vocabularies.

Figure 3 include three main entities, namely **Object**, **Annotation** and **User**. *Objects* are the prime input for the system. They refer to each collection object with a unique identifier and represent basic registration properties (these could differ by type and number from institution to institution). For example, as we saw in Section 2, the RMA registers each object at least with its identifier, title and dimensions. Additionally, it could provide a basic or detailed *Annotation* including a short description, associated people and roles, materials, techniques and iconography references (using Iconclass terms). An annotation also includes a reference to the *User* (internal or external to the museum) who created it, and the date of creation. Each user can have different *Roles*, such as cataloger, crowd annotator, curator or administrator. A user is typically associated with a login and basic properties like email-address and name.

To support the creation of annotations we introduce additional entities - **Vocabulary**, **Topic** and **Document** (Figure 4). A *Vocabulary* typically refers to a thesaurus, taxonomy or other structured set of terms (*Topics*), for example Iconclass, ULAN[10]/VIAF, AAT[11], GeoNames[12]. Each *Vocabulary* can provide a different set of descriptive properties for their terms. For example, in Iconclass terms have an identifier, a description and relations to other terms such as narrower, broader and related. As described in Section 2, each annotation refers to one or more *Documents* (sources), e.g. web pages or books providing detailed information on the main topics in the annotation. URLs and titles are used as references in each annotation.

To support updating of annotations and different roles of users we introduce the **Review** entity (Figure 4), which resembles the current quality state of an *Annotation* given by a *User* with a curator role.

5.2 Components

For the system to support the four steps in the workflow (*identify object to annotate*, *assign objects to annotate*, *support annotation of objects* and *assess quality of objects*) we identified methods to perform the functions of the steps, clustered the methods into logical components, and identified components to interact with the stored data (Figure 4): the resulting component diagram is shown in Figure 5. We make a distinc-

---

5 http://purl.org/dc/terms/
6 http://xmlns.com/foaf/0.1/
7 http://www.openannotation.org/ns/
8 http://purl.org/stuff/rev#
9 http://www.w3.org/2004/02/skos/core#
10 http://www.getty.edu/research/tools/vocabularies/ulan/index.html
11 http://www.getty.edu/research/tools/vocabularies/aat/index.html
12 http://www.geonames.org/

tion between components which are able to store and manipulate stored data (in yellow) and components which use the data (in orange).

The four main components (in orange) directly map to the workflow in Figure 2 and use the data provided by the data components. The **Identify** component finds subsets of the collection data that are suitable to include in the annotation task. The **Assign** component matches users with specific Object that follow from the **Identify** component. The annotation task is supported by the **Support** component, which includes two data sub-components for recommending Topics and Documents. The **Assess** component supports the assessment of the quality of Annotations done by Users.

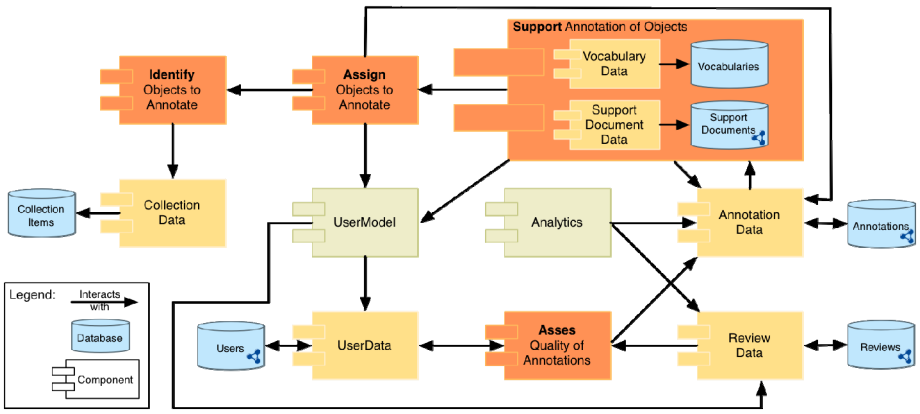

Fig. 4. Software component diagram

In total there are six data components: **Collection Data**, **Annotation Data**, **User Data**, **Vocabulary Data**, **Support Document Data** and **Review Data**. The Collection Data and Vocabularies Data components contain methods that can convert the datasets for Collection Items and Vocabularies into RDF if that has to be done (these datasets are maintained by the CH institution itself). Each data component supports basic query operations in addition to methods enabling convenient access to frequently used data.

Reviews and Annotations are used by the **Analytics** component to provide information about the Users and the system performance. Information from this component is used in the diagnosis and evolution of the system. The **User Model** component uses annotation and user data to determine relevant aggregate values for Users such as expertise and interest on Topics.

We implement these components with the possibility to vary between different strategies to implement the functionality. We list some strategies related to personalization:

- Component Assign - AnnotatorSimilarity, AnnotatorExpertise, Random
- Component Vocabulary – CommandOfVocabulary
- Component SupportDocuments – DocumentAccess
- Component Assess - ReviewerExpertise, Random

For example, for matching annotators to items the Assign component could be implemented using an approach based on what similar annotators did, the interest of the annotator, the expertise of the annotator, or provide the annotator with a random item.

5.3 Generic approach to Strategies

For the purpose of applying the new system in an explorative way, we have chosen to make the input parameters and strategies for each compnent configurable. A minimal number of interventions is thus needed to reconfigure the system. It is also possible to easily combine different strategies. Since this is an important aspect of our approach, especially to tune the personalization support, we use interfaces so that new strategies can be created without having to change the implementation of other parts of the system.

Each configurable piece in a component allows for different strategies to be chosen, but these strategies should match in the method names and return types. For our proposed system we created the interface `IStrategy` with two method signatures. The `SetParameter` method enables the system to set an arbitrary number of parameters that a strategy needs. The `GetResults` method executes the strategy and returns the result. The `<T>` construction allows methods to operate on objects of various types. The parameter `previousResult` allows strategies to be combined so that the output of one strategy is the input of the other strategy.

```
Interface IStrategy<T>{
  void SetParameters(Map<String,Object> params);
  T GetResults(T prevResult); }
```

5.4 Strategy Implementation

The **Assign** component assigns *Items* to a *User*. The strategy to select which *Items* to assign is configurable such that the strategy is a parameter of the function that returns list of *Items*. We provide pseudo code for a strategy that is based on the User's expertise. Non-essential parts are removed.

```
class ExpertiseStrategy implements IStrategy<List<Item>>{
  Map<Topic,double> exp; //Topical expertise of the user
  double threshold; //Threshold for expertise on topics
  void SetParameters(Map<String,Object> params){
    exp = params["exp"] as Map<Topic,double>;
    threshold = params["threshold"] as int; }
  List<Item> GetResults(List<Item> prevResult){
   List<Item> result = new List<Item>();
     foreach(Item i: prevResult){
       foreach(Topic t: i.Topics){
         if(exp[t] > threshold){
           result.add(i); break; }}}
  return result; }
```

With this strategy implemented the **Assign** component can be configured to use this strategy. The method that calls the **Assign** component should then make sure the correct parameters are set. The input data (`prevResult`) can be retrieved by calling the **Identify** component. The strategy used in the **Identify** component can also be configured. Such a strategy also implements `IStrategy` but possibly with a different `<T>`.

6 Evolutionary Design Process

The research method that we employ in this project implies that our proposed system will be used to explore different ways of supporting crowd-based annotation. Our system is therefore customizable to allow for easy changes in input parameters and used strategies. In Section 6.1 we mention metrics we will use to measure the performance of the system and in Section 6.2 we explain how this information is applied.

6.1 Metrics

In addition to supporting the crowd-based annotation, we also measure the performance of the system in doing so. This is done in the **Analytics** component. These measurements allow us to understand and subsequently tune the effect of allowing external annotators in the process. The component interacts with the **Annotation Data** component for quantitative data and with **Review Data** for qualitative data.

Quantitative metrics indicate the crowd annotation process performance and include for example the number of annotations done in total, on average per user, on average per object, and on average per topic. These metrics are also available for *Reviews* and per specific *Topic*.

Qualitative metrics indicate the quality of the crowd-based annotations, based on the Reviews done. This helps the RMA and us in understanding the effect of the configured personalized strategies. Metrics include the overall percentage of correct annotations, per *User* and/or per *Topic*. Other qualitative metrics help in understanding whether the provided supporting documents are indeed used and whether that increases the quality of the annotations.

6.2 Dashboards

A personalized dashboard provides each individual *U*ser with insights into what they annotated, the *Topic*s they annotated, and how well they annotated. The estimated expertise level for *Topics* and the *Topics* that need improvement are shown. In that dashboard users can also indicate that they have a higher or lower interest for certain *Topics*. Next to a personalized dashboard for each individual *User*, the system has a global dashboard, with information about all *Users*. Part of that shows general information, like the quantitative metrics, to visitors and *Users* of the system; the rest of it helps RMA and us in diagnosing and tuning the configurations of the crowd annotation process.

7 Conclusion

We have analyzed an object digitization and annotation workflow representative for Cultural Heritage institutions. Based on this analysis we identified four tasks of the workflow which can benefit from involving people from the crowd: identifying objects to annotate, assigning objects to annotate, supporting annotation of objects, and assessing quality of objects. We designed a customizable system to support crowd-based annotation where these four tasks are central elements.

The next step will be to fully implement the described system. We will further develop the personalization strategies and explore which of them work well in supporting the crowd in the annotation task. Challenges in this exploration will include maintaining the high level of quality of annotations, addressing the appropriate persons within the crowd and keeping them involved.

Acknowledgements. This publication was supported by the Dutch national program COMMIT.

References

1. Hyvönen, E., Mäkelä, E., Salminen, M., Valo, A., Viljanen, K., Saarela, S., Junnila, M., Kettula, S.: MuseumFinland—Finnish museums on the semantic web, Web Semantics: Science, Services and Agents on the World Wide Web 3 (2005) pp. 224-241
2. Wang, Y., Stash, N., Aroyo, L., Gorgels, P., Rutledge, L., Schreiber, G.: Recommendations based on semantically enriched museum collections. Web Semantics: Science, Services and Agents on the World Wide Web 6 (2008) pp. 283-290
3. Oomen, J., Aroyo, L.: Crowdsourcing in the cultural heritage domain: opportunities and challenges. In: Proceedings of the 5th International Conference on Communities and Technologies, pp. 138–149. ACM, New York (2011)
4. Cosley, D., Frankowski, D., Terveen, L., Riedl, J.: SuggestBot. In: Proceedings of the 12th international conference on Intelligent user interfaces, pp. 32-41. ACM, New York (2007)
5. Hildebrand, M., Van Ossenbruggen, J., Hardman, L., Jacobs, G.: Supporting subject matter annotation using heterogeneous thesauri: A user study in Web data reuse. International Journal of Human-Computer Studies 67 (2009) 887-902