Human Computation (2019) 6:1:12-41

ISSN: 2330-8001, DOI: 10.15346/hc.v6i1.2

Accurator: Nichesourcing for Cultural Heritage

CHRIS DIJKSHOORN, VRIJE UNIVERSITEIT AMSTERDAM

VICTOR DE BOER, VRIJE UNIVERSITEIT AMSTERDAM

LORA AROYO, VRIJE UNIVERSITEIT AMSTERDAM

GUUS SCHREIBER, VRIJE UNIVERSITEIT AMSTERDAM

ABSTRACT

With the increase of cultural heritage data published online, the usefulness of data hinges on the quality and diversity of descriptions of collection objects. In many cases, existing descriptions are not sufficient for retrieval and research tasks, resulting in the need for more specific annotations. Eliciting such annotations is a challenge since it often requires domain-specific knowledge. Where crowdsourcing can be successfully used to execute simple annotation tasks, identifying people with the required expertise might prove troublesome for more complex and domain-specific tasks. Nichesourcing addresses this problem, by tapping into the expert knowledge available in niche communities. This paper presents Accurator, a methodology for conducting nichesourcing campaigns, by addressing communities, organizing events and tailoring a web-based annotation tool to a domain of choice. The contribution of this paper is fourfold: 1) a nichesourcing methodology, 2) an annotation tool for experts, 3) validation of the methodology in three case studies and 4) a dataset including the obtained annotations. The three domains of the case studies are birds on art, bible prints and fashion images. We compare the quality and quantity of obtained annotations in the three case studies, showing that the nichesourcing methodology in combination with the image annotation tool can be used to collect high-quality annotations in a variety of domains. A user evaluation indicates the tool is suited and usable for domain-specific annotation tasks.

1. INTRODUCTION

Many cultural heritage collections are currently being made available online (Mouromtsev et al., 2015; Szekely et al., 2013; de Boer et al., 2012b). While such online collections can be valuable resources for the general public, scholars and professional users, their usefulness depends on correct and rich descriptions of the contained objects. Metadata describing objects is usually created by professionals working for the cultural heritage institution and typically meets the needs of other cultural heritage professionals. Many institutions lack the manpower to adapt data in order to better support different groups of users. Therefore, some institutions have turned to crowdsourcing,

outsourcing tasks to a distributed and often anonymous group of people (Oomen and Aroyo, 2011). For cultural heritage organizations, crowdsourcing proved to be a low-cost solution to gather large quantities of descriptions (Chun et al., 2006; Ellis et al., 2012; Gligorov et al., 2013).

While many institutions have gained significant experience with using crowdsourcing to collect large quantities of data, a remaining challenge is how to best harness the diversity in the crowd to solve difficult tasks in a sustainable fashion (Noordegraaf et al., 2014). Describing collection objects is a knowledge-intensive task, due to the variation in types of objects, diversity in subject matter and sometimes hidden symbolic meaning. Accurately annotating objects therefore often requires domain-specific knowledge. At the moment the required expertise is unavailable in an organization and when it is unfeasible to hire professionals to do the work, it is fruitful to reach out to experts within the crowd.

Nichesourcing is a type of crowdsourcing, where groups of people with domain-specific knowledge are involved in the annotation process (de Boer et al., 2012a; Dijkshoorn et al., 2013). We call these groups of enthusiasts niche communities. There are numerous niche communities out there, revolving around lots of different domains. The advantages of nichesourcing are: 1) contributors are intrinsically motivated, 2) there is the potential of obtaining annotations of higher quality and 3) knowledge-intensive annotation tasks can be executed.

Where de Boer et al. (2012a) introduced the idea of nichesourcing and discussed small-scale case studies, a structured methodology was missing. We here present a repeatable and sustainable methodology as well as an open-source tool to support nichesourcing. We validate both using three extensive real-world case studies. The contribution of this paper is fourfold:

- **Accurator nichesourcing methodology** which provides a step-by-step guide to designing and executing a nichesourcing campaign (Section 3)
- **Accurator annotation tool** that supports the nichesourcing process (Section 4)
- **Validation of nichesourcing methodology** in three case studies in different domains (Section 5)
- **Dataset of annotations** which includes the annotations obtained during the three campaigns (Section 6)

Section 6 includes an analysis of the annotations and an evaluation of the annotation tool. The paper is concluded with a discussion and future work section.

2. RELATED WORK

Human computation is a field in which the human ability to carry out computational tasks is leveraged to solve problems that can not yet be solved by computers alone (Quinn and Bederson, 2011). Crowdsourcing is part of the human computation field and regards tasks that are outsourced to a large group of people, often using the internet as an intermediary (Doan et al., 2011). Crowdsourcing proved to be a good way to gather annotations at scale (von Ahn and Dabbish, 2004; Raddick et al., 2010). This was recognized by the cultural heritage community and crowdsourcing has been used to annotate objects such as paintings, maps and videos (Ellis et al., 2012; Simon et al., 2011; Gligorov et al., 2011). Gathered annotations are complementary to annotations provided by cultural heritage professionals and thereby improved the accessibility of collections (Chun et al., 2006;

Gligorov et al., 2013). Crowdsourcing turned out to be a novel way of engaging the public as well (Ridge, 2013).

Despite many successes, some crowdsourcing projects fail to live up to their expectations. Research has been conducted in classifying different types of crowdsourcing initiatives, to predict their success based on project characteristics (Noordegraaf et al., 2014). *Methodology papers* that outline steps to successfully run a crowdsourcing campaign are however scarce. Yadav and Darlington (2016) discuss guidelines to how Semantic Web technology can support the design and management of crowdsourcing projects, while Sarasua et al. (2015) introduce guidelines for designing platforms hosting multiple projects. In this paper we specify a nichesourcing methodology, contributing to the work available on crowdsourcing methodologies. More specifically, the methodology addresses crowdsourcing challenges such as solving knowledge-intensive tasks, involving experts, motivating contributors and assuring high-quality contributions.

Different approaches have been proposed to *solving knowledge-intensive, domain-specific tasks*. von Ahn and Dabbish (2004) introduce theme rooms, clustering tasks by domain and leaving the choice for a task to the contributor. Finding tasks can also be automated: task assignment matches characteristics of contributors with suitable tasks (Cosley et al., 2007; Difallah et al., 2013). Kulkarni et al. (2014) search for experts in the crowd to improve complex, creative tasks. Combinations of improvement tasks can be optimized in crowdsourcing workflows, by considering the average ability of contributors, the variance in the ability of contributors and improvement difficulty (Goto et al., 2016). Oosterman and Houben (2016) invite experts from online communities to annotate objects in a specific domain. Ipeirotis and Gabrilovich (2014) use online advertisement platforms in combination with quizzes to target knowledgeable contributors. A different approach is to teach contributors how to solve knowledge-intensive tasks using a game (Traub et al., 2014). Chamberlain (2014) investigates the ability of groups on social networks to solve tasks, concluding that topic-specific groups are more active and solve more tasks. Nichesourcing builds upon these approaches by involving off- and online niche communities to solve knowledge-intensive tasks.

3. ACCURATOR NICHESOURCING METHODOLOGY

In this section, we describe the Accurator nichesourcing methodology. Figure 1 provides a schematic overview of the methodology, which consists of four stages: orientation, implementation, execution and evaluation. The methodology is cyclic, one iteration can build upon the results obtained during a previous iteration. The stages are further segmented into steps. While omitting some steps could still yield valuable results, we argue in this section why it is important to include all steps in the application of the methodology, in order to conduct successful nichesourcing campaigns. In this section, we describe for each of these steps the input, output, action and challenges. We start with a definition of nichesourcing and an introduction to the terminology used in this paper.

Based upon the paper by de Boer et al. (2012a) we provide the following definition for nichesourcing:

> *Nichesourcing: the practice of completing knowledge-intensive tasks, by soliciting niche communities with the required domain-specific knowledge.*

Nichesourcing extends crowdsourcing in the sense that rather than executing simple micro-tasks, domain-specific, knowledge-intensive tasks can be executed by intrinsically motivated members of

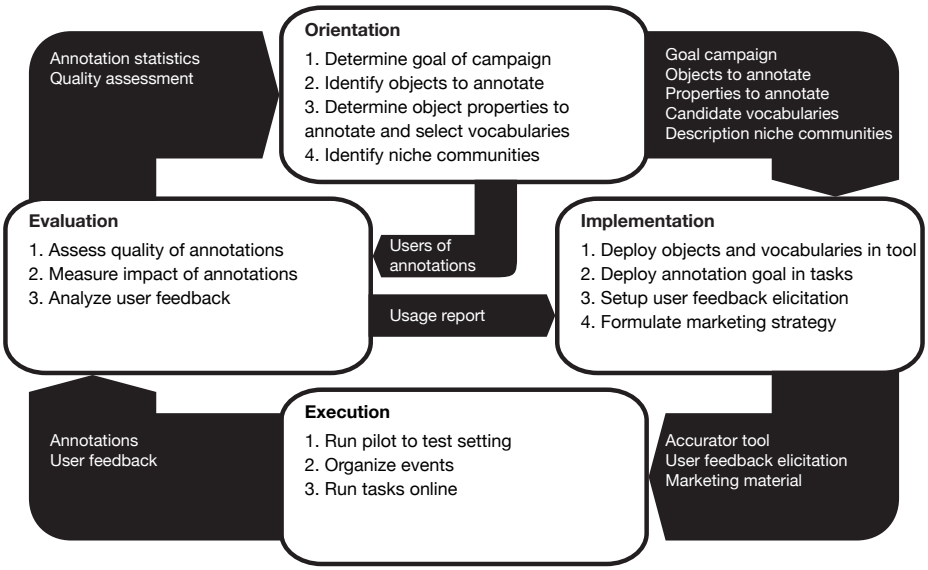


Figure 1. The four stages of the Accurator nichesourcing methodology.

communities, able to provide high-quality results. *Niche communities* have a shared domain of interest, members have an affinity with this domain and their regular interactions engenders social trust and reputation. These niche communities can correspond to the notion of a community of practice or interest (Wenger et al., 2002). Examples of *domains* are ornithology and fashion. We continue by listing the terminology relevant to the nichesourcing methodology:

- **requesters** initiate nichesourcing campaigns and often correspond with the cultural heritage institution that owns the collection objects.
- **collection objects** are real-world objects such as paintings or prints, of which images can be used in online applications.
- **annotations** are often short textual descriptions or concepts, that can be used to describe images of collection objects.
- **tasks** combine collection objects with the sort of annotations requested.
- **contributors** solve tasks. We refrain from using the denomination worker since there is no monetary reward given for completing tasks.
- **task difficulty** indicates how hard it is to solve a task.
- **contributor ability** is an indication of how well a contributor can solve hard tasks.

The terminology above will be used throughout the description of the different stages of the Accurator nichesourcing methodology.

3.1. Orientation stage

As shown in Figure 1, in the first stage, the goal of the campaign is determined and the objects that need to be annotated to reach this goal are identified. Based on the characteristics of these objects,

the required enrichment is determined, which guides the identification of niche communities who can provide such information.

1. Determine goal of campaign

input: annotation statistics *output:* goal campaign, users of annotations

action: The requester formulates a goal in this first step, stating what an institution wants to achieve with the annotations obtained during a nichesourcing campaign. The formulated goal can range from general (e.g. to improve access to a collection) to specific (e.g. to answer a digital humanities research question). A goal is based either on needs internal or external to the institution, therefore it is accompanied by an overview of the intended users of the annotations. A list of users provides clarity about who benefits from the data and gives an indication of who will take care of the collected information when the campaign is completed. At the end of the campaign, the goal is used to verify whether the gathered annotations have the desired impact. If the annotation statistics indicate the goal is reached, a new goal is formulated, otherwise, a subsequent improved campaign is used to reach the current goal.

challenges: The goal has to be formulated in such a way that contributors deem it worthy to invest time into, fitting with their domain of interest. Additionally, it helps to validate the results of a campaign if it is possible to measure whether a goal is reached. For example, if the aim is to improve access to a collection, this can be measured by standard information retrieval metrics such as precision and recall.

2. Identify objects to annotate

input: goal campaign *output:* objects to annotate

action: During this step, a subset of objects is identified, which when correctly annotated will bring us closer to the formulated goal. This can be a) on the basis of automatic (data-driven) analysis; b) through manual selection of objects that require improvement or c) by analyzing user interactions with the collection. Most cultural heritage collections consist of objects that relate to a range of different domains. To be suitable for nichesourcing, the selected objects should share a domain, which later is matched with a community of experts.

challenges: Once a set of collection objects is either automatically or manually identified, preparation steps might be needed to ensure that basic metadata and an image are available for every object in the set. Additionally, intellectual property rights should allow images of the objects to be used online.

3. Determine object properties to annotate and select vocabularies

input: objects to annotate *output:* properties to annotate, candidate vocabularies

action: The object properties that need to be annotated to achieve the goal are identified during this step. More specifically, we distinguish properties of the objects that can be better described using numerical values, textual descriptions or concepts from structured vocabularies. One specific goal here is to identify structured vocabularies that can be used as values for the annotations. These vocabularies can be provided as input to the Accurator annotation tool, which presents concepts of the vocabulary as options to the contributors.

challenges: Cultural heritage institutions have to carefully consider which vocabularies to use for

describing collection objects. The suitability of vocabularies should be assessed in terms of completeness, accuracy and original context. It can, for example, be that the vocabulary was intended to be used in a completely different context and therefore does not contain the desired concepts, or that concepts represent a worldview which is different from the institution. A lack of available labels in some language can pose a more direct problem.

4. Identify niche communities

input: quality assessment *output:* description niche communities

action: To assess the feasibility of nichesourcing, the shared domain of the set of objects should match with a niche community. These communities are identified by contemplating on which people have the expertise to annotate the objects. The characteristics of an object can for example match with professionals outside the cultural heritage sector, or with hobbyists focusing on a certain topic. A common feature of niche communities is that they can be divided into even more specialized sub-niches. It is useful to identify such sub-niches, since later in the process it helps to assign tasks to contributors most knowledgeable of a sub-niche.

challenges: The description of niche communities should include ways of reaching out to the community, which is important for the marketing strategy in the implementation stage. Furthermore, the niche community should not only be determined on the basis of the match with the objects but more importantly, on the match with the missing information. It is not always straightforward to identify niche communities that match the selected objects and requested information. It is therefore important to allow interplay between the steps, adapting the selection or requested information to the communities available.

3.2. Implementation stage

In the second stage of the methodology shown in Figure 1, the Accurator annotation tool is deployed and tasks are designed that help reach the goal of the campaign. A marketing strategy is formulated to address the niche communities.

1. Deploy objects and vocabularies in tool

input: objects to annotate, candidate vocabularies *output:* accurator annotation tool

action: In this step, the Accurator annotation tool is deployed and relevant data is loaded. We describe the tool in more detail in Section 4, but in general, this requires a requester to a) set up a server environment; b) install the tool and c) adapt the tool to the domain. Once deployed, data regarding the selected objects and vocabularies is loaded. A single instance of a tool can accommodate multiple campaigns, refraining an institution from having to deploy a tool for each iteration of the methodology.

challenges: Deploying the Accurator annotation tool requires technical knowledge as well as appropriate infrastructure. Not every institution will have both readily available and therefore some might choose to outsource this step. Alternatively, an institution can choose to use existing online crowdsourcing platforms (e.g. Amazon Turk), thereby bypassing this problem. This has the downside that these platforms are not easily customized to support a particular domain.

2. Deploy annotation goal in tasks

input: goal campaign *output:* accurator annotation tool

action: During this step, the goal of the campaign is translated into smaller annotation tasks. Tasks combine objects with explicit requests for information and instructions on how this information should be provided. For a photograph, the requested information could be depicted persons, accompanied by the instruction to enter names into a text field. The identified structured vocabularies are related to requests, allowing rendering of suggestions for values to enter. Tasks are defined in the annotation tool by relating the identified objects to input fields, each accompanied by the information request and structured vocabulary.

challenges: The request for information and instructions have to be concise and unambiguous. If there is room for interpretation, this will have a negative impact on the consistency of the provided annotations. The concepts suggested can help normalize the input, but should fit the type of information requested.

3. Set up user feedback elicitation

input: - output: user feedback elicitation

action: To get insights into the behavior of users and collect feedback, user elicitation mechanisms are set up (Albert and Tullis, 2013). These mechanisms can be automated and unobtrusive, such as logging interactions with the annotation tool. An institution can also choose for more direct inquiring, for example, by using questionnaires. Information gathered using these mechanisms is used to refine the orientation stage and can indicate the effectiveness of a marketing strategy. Furthermore, created user profiles can serve as input for automated quality assessment of annotations (Ceolin et al., 2012).

challenges: Nichesourcing relies on the intrinsic motivation of contributors. To not annoy contributors and distract them from solving tasks, the elicitation mechanisms should be as unobtrusive as possible.

4. Formulate marketing strategy

input: description niche communities *output:* marketing material and schedule

action: A marketing strategy is formulated to engage niche communities and capture the attention of contributors (Palmatier and Sridhar, 2017). This strategy includes a schedule that details when and how messages are communicated. Different outlets can be used, such as social media, newsletters and flyers. The choice of outlet depends on how the targeted niche community can best be reached. First communications are focussed on drawing attention to the campaign, by inviting people to participate in annotation events. Systematic quality feedback helps to retain contributors and subsequent communications are meant to entice people to keep participating in campaigns (Moon and Sproull, 2008). Following an event, a message can be sent about the progress made, in addition to an invitation to keep contributing online. At the end of the campaign, the impact of the annotations is emphasized, alongside pointing contributors towards new campaigns when available.

challenges: It can be challenging to reach the niche communities identified during the orientation stage. Sometimes organizations that already rally events around the domain of interest can serve as a point of entry. These organizations are often different from the cultural heritage institution that owns the collection. Finding a niche representative within such an organization, who is willing to collaborate, greatly eases addressing potential contributors. Another strategy is to market the

nichesourcing campaign together with a broader event associated with the domain, for example, a National Week of Fashion or an exhibition organized by the institution. This allows institutions to combine the effort needed for marketing.

3.3. Execution stage

With the tool deployed and the marketing strategy in place, the nichesourcing tasks can be executed (Figure 1). But first, tasks deployed in the annotation tool are tested during a pilot.

1. Run pilot to test the setting

input: accurator annotation tool *output:* -

action: To test the annotation tool and formulated tasks, a pilot is run with a limited number of members of the targeted niche community. During the pilot, issues are identified that should be addressed before the event. Depending on the type of issue, the subset of objects, selected vocabularies and tasks are refined.

challenges: For each issue, an assessment has to be made whether it will apply to most members of the community and therefore warrants a follow-up action.

2. Organize events

input: accurator annotation tool, user feedback elicitation, marketing material, schedule *output:* annotations, user feedback

action: Organizing an annotation event is an essential element of an Accurator nichesourcing campaign. Besides being the first source of annotations and feedback, the event is used to engage the niche community. The organization of events constitutes of three aspects: timing, location and program. With respect to timing, enough time is needed to implement the marketing strategy and advertise the event in the niche community. To make the event as attractive as possible, the event should preferably take place at a location relevant to the domain of interest. This could be at the institutions of the collection owner, or at another place relevant to the domain. The program of the event includes an introduction and demonstration of the tool. After this, contributors use the tool to annotate the collection objects. The event is concluded with a discussion, resulting in feedback which can be used during the evaluation stage. Optionally, the program can be extended with additional activities, functioning as an incentive for experts to participate.

challenges: It can be challenging to strike the right balance between time for annotating, discussion and extra activities. Enough time has to be available for annotating collection objects, in order to collect sizable amounts of annotations and to make sure that contributors have enough time to work with the tool to be able to provide feedback.

3. Run tasks online

input: accurator annotation tool, user feedback elicitation, marketing material, schedule *output:* annotations, user feedback

action: Following an annotation event, the campaign is continued online. Running the nichesourcing tasks online regards advertising the annotation tool and providing support to contributors. The interest sparked up by the event serves as initial input for advertising the tool. Updating contributors on the results of the annotation event helps to incentivize people to return at a later point in time and

continuously add annotations to the collection. To sustain this attention and reach new contributors, the tool is advertised as outlined in the marketing campaign. Finding additional experts could be automated using techniques such as proposed by Kulkarni et al. (2014); Ipeirotis and Gabrilovich (2014) and Oosterman and Houben (2016). In order for contributors to not get discouraged when they run into problems, adequate support has to be available.

challenges: To sustain the interest of contributors, a cultural heritage institution will have to invest in the support and marketing of the annotation tool. When a group of contributors is actively involved in the nichesourcing campaign, the effort of marketing and providing support can be shifted towards the community (Bevan et al., 2014).

3.4. Evaluation stage

At the end of the nichesourcing campaign, the impact and quality of the annotations are assessed. As shown in Figure 1, feedback gathered during the campaign is used to improve subsequent campaigns.

1. Assess quality of annotations

input: annotations *output:* quality assessment

action: The quality of annotations is assessed during this step. Quality verification procedures can be manual processes or automated processes. Both can be used within a nichesourcing campaign, although their suitability should be assessed up front. An example of a manual process is reviewing (parts of) the annotations, by contributors or professionals. Probabilistic methods can, for example, be used to automate the assessment process (Whitehill et al., 2009). An institution decides based on the assessment, to reject or improve annotations (Goto et al., 2016). Institutions should consider publishing the annotations along with their quality assessment since further analysis of measured disagreement can lead to new insights in crowdsourced data (Inel et al., 2014).

challenges: A relatively naive automated method such as majority voting might be less appropriate for nichesourcing since a small number of experts might be knowledgeable enough to provide a correct annotation. An annotation which might, in turn, contradict annotations of other contributors. Other automated approaches would, therefore, be more suitable, for example considering trust in a contributor based on earlier annotations (Ceolin et al., 2012).

2. Measure impact of annotations

input: users of annotations, annotations *output:* annotation statistics

action: During this step, the verified annotations are deployed to investigate whether the goal is reached. If the goal is to improve accessibility and the user of the annotations is the institution, this, for example, entails exporting the data from the tool and incorporating the results into the collection data. At that point, a comparison of search performance of the collection with and without the annotations can provide an indication of impact (Gligorov et al., 2013). If the goal cannot be reached, this evaluation serves as input for improving the next nichesourcing campaign, by for example adapting the set of objects or the properties to annotate.

challenges: Quantifying the impact of annotations can be difficult and depends on the formulated goal. Thereafter, it can be challenging to translate this evaluation towards adaptions of the next nichesourcing campaign.

3. Analyze user feedback

input: user feedback *output:* usage report

action: Feedback is gathered during events as well as online. User feedback follows from sources such as questionnaires, discussions, support requests and interaction logs. Analyzing these sources can help to improve subsequent nichesourcing campaigns. Common feedback topics regard task complexity and appropriateness of the tool. If tasks are deemed too complex, changes can be made to the selection of objects, chosen properties to annotate and the niche community which is addressed. When tasks are too easy, other crowdsourcing approaches could be considered. Feedback regarding the tool can be addressed by improving the code or choosing a different platform to deploy tasks.

challenges: Operationalizing the gathered feedback, by improving new campaigns, can be a challenge. It is, however, important to acknowledge feedback and improve the process. A contributor providing feedback took the time to work with the tool and provide feedback. If this feedback is taken seriously, a contributor might feel more inclined to contribute to a new campaign. Addressing problems with tooling requires technical skills which might not be available within an institution. The shortcomings of a tool could, therefore, be communicated to the contributors providing feedback, or programmers could be contacted to improve the tool. The Accurator annotation tool, which we discuss in the next section is open source, allowing anyone to improve the code as desired.

4. ACCURATOR ANNOTATION TOOL

To support the nichesourcing methodology, we present a tool called *Accurator*. Accurator is a web-based annotation tool which can be instantiated for specific nichesourcing campaigns, to allow contributors to annotate images of cultural heritage objects that are automatically assigned to them. The adaptability to a specific domain and intrinsic use of interoperable data makes the tool different from other annotation tools. This section describes the tool and more specifically its adaptability to domains, task assignment approaches, as well as usability design considerations. We conclude this section by discussing how the collected data can be used by other systems and how the annotations directly impact the search functionality of the tool.

The implementation of the tool is based on Semantic Web technology (Shadbolt et al., 2006). The Accurator annotation tool is available as a package for the Cliopatria Semantic Web infrastructure (Wielemaker et al., 2016). The back-end is written in the Prolog programming language, which facilitates direct access to the data layer (Wielemaker et al., 2007). The front-end uses jQuery[1] and Twitter Bootstrap[2] so contributors experience an interactive and responsive tool. The source code of the package is published online, along with an in-depth guide to how new instances can be deployed[3].

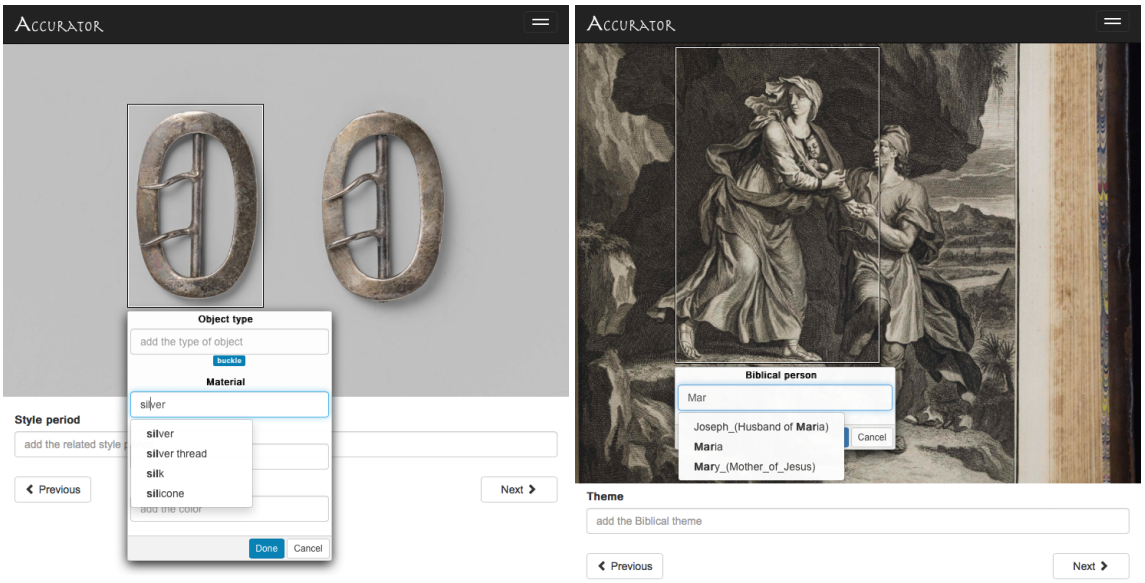
4.1. Adaptability to the domain

Accurator can be customized to fit a domain, by using config files containing *domain definitions*. The definitions define links to 1) a specification of the annotation fields relevant to the domain, 2)

[1] https://jquery.com

[2] https://getbootstrap.com

[3] https://github.com/rasvaan/accurator



(a) Fashion domain *(b) Bible domain*

***Figure 2.** Annotation interface of the Accurator tool, showing the fields that can be used to annotate objects in the fashion domain and bible domain.*

elements of the interface tailored to the domain and 3) information that enables task assignment. Here we discuss annotation fields and interface adaption, task assignment follows in a separate subsection.

Annotation tasks are adaptable to the domain, a requester can specify field definitions for each of the annotation fields. These specifications include the field name, a short instruction and the type of field. Different types include radio buttons, check boxes and text fields. Text fields can use the auto-completion functionality, where a contributor starts to type and a drop-down menu renders alternatives related to this input, as shown in Figure 2. The contributor can either choose to annotate the object using one of these alternatives or use the entered text. The alternatives originate either from a list of values added to the field definition or from a subset of a structured vocabulary. Accurator includes Prolog predicates to identify such subsets of vocabularies, for example, based on a branch within a taxonomy.

Annotation fields can be defined as being about the object as a whole, or be defined as being about a specific part of the object. In the first case, annotation fields are presented to a contributor alongside the image. An example of this is the style period of fashion objects. In the second case, users can draw a bounding box in the image to identify the specific part of the object that the annotation concerns, as shown in Figure 2. This allows users to annotate multiple specific elements of an object, for example, two birds of different species depicted on a print.

The default visual elements and text of the tool can be adapted as well, the default tagline used on the intro page of the tool “Help us add information to artworks” can, for example, be changed to

one tailored to the fashion domain (e.g. "Help us describe fashion"). At the same time, it is possible to add images, which brand the tool with visuals related to the domain.

4.2. Task assignment

Task assignment concerns the matching of contributors with tasks. Accurator provides three modes of task assignment: *ranked*, *sub-domain based* and *recommendation*. Ranked is the default mode, which first filters out the objects already annotated by the user and sorts the remaining objects based on the total number of users that annotated them. A list of objects randomly picked from the least annotated objects is presented to the contributor. This is the default setting since it ensures a rapid increase in annotated objects.

In the sub-domain based mode, contributors can choose in which sub-domain they would like to annotate. To this end, a hierarchy of general and more specific domains is created, by adding references to sub-domain definitions in the configuration file of a domain definition. The fashion domain described in Section 5.3 can, for example, be split into more specific domains such as costumes and jewelry. The availability of sub-domains triggers a finer grained mode of task assignment, as the objects presented to contributors are filtered based on the domain they belong to. The objects from the domain chosen by the contributor are then ranked according to the ranked method described above.

The third mode of task assignment is recommendation. Recommending suitable tasks to contributors might make the annotation process more accurate and efficient. With the Accurator tool, we experimented with recommendation based on the elicitation of expertise levels of contributors. To do this, a list of expertise topics is created, the expertise levels from contributors are elicited and the obtained levels are used as input for a recommender algorithm. The list of topics is based on a structured vocabulary, referenced in the configuration file. In case of the birds on art domain, an example of topics can be a branch of the biological taxonomy. Contributors are asked to assess their expertise regarding each selected topic. The highest ranking topics are used as input for an explorative search algorithm, which uses the graph structure to find objects that are related to the expertise of the contributor (Wielemaker et al., 2008). In Section 6, we evaluate the three different task assignment approaches and consider the feedback of contributors.

4.3. Usability

Usability is important for crowdsourcing tools, and we argue that this is especially true for tools that are used for nichesourcing since nichesourcing relies on the intrinsic motivation of contributors. Wasting their goodwill because the tool is hard to use might make a requester miss out on valuable input. While the tool uses many Semantic Web techniques, as outlined by Sarasua et al. (2015), we should not expect our contributors to be Semantic Web experts. The interface, therefore, hides technical aspects such as the persistent identifiers from contributors and uses textual labels of concepts and properties whenever available.

Part of the usability is presenting a tool in the language of the contributor. The primary language of the annotation tool is English, but many of the contributors prefer a different language. The tool supports translating textual elements of the interface, in a similar fashion as adapting texts to the domain. We translated the interface to Dutch, thereby customizing the usage for contributors from

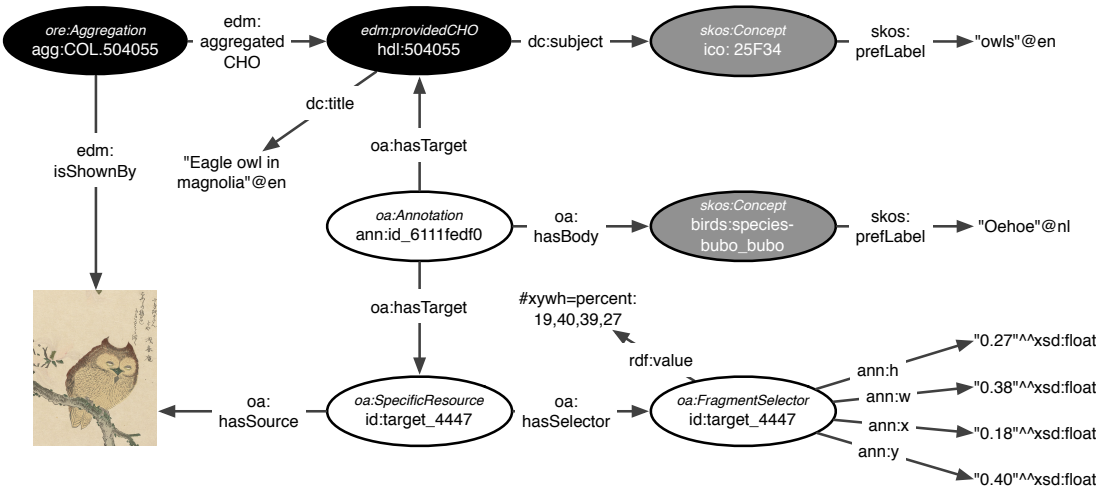


***Figure 3.** A graph representation of the print "Eagle owl in magnolia" annotated with the species of the depicted bird.*

the Netherlands. The auto-completion alternatives are based on the labels of concepts of structured vocabularies. Oftentimes these labels are available in multiple languages. The tool is designed to render alternatives in the language of choice if available, otherwise falling back on English labels.

The Accurator annotation tool is designed to work with all regular browsers, even older versions. Therefore most contributors will be able to use the tool on their own system. The registration procedure is simple and requires minimal information to be entered by potential contributors. Questions requesting additional information about contributors used for scientific purposes are spread out over multiple blocks, each appearing after a contributor added a specified number of annotations. Additionally, system administrators are advised to use simple domain names for the online tool, so contributors can easily remember how to access the instance. We evaluated how contributors perceive the usability of the tool using a questionnaire, the results are discussed in Section 6.

4.4. Direct impact annotations

Annotations added by contributors can be directly used by other systems and have a direct impact on the semantic search functionality of the Accurator tool. Annotation data is stored in the triple store of the annotation tool, which is separated from the collection management system or catalog of the institution. Using this architecture, systems that cope well with crowdsourced data can have direct access to new information, while systems relying on verified data can use exports of the information of which the quality is assessed. Storing the data using the Resource Description Framework (RDF)[4] and standardized data models improves the reusability of data. We continue by discussing the data model used within the tool, followed by a discussion of the impact of annotations on search and the ways of how the collected data can be made available.

[4]https://www.w3.org/RDF/

A graph representation of information describing a print of the Rijksmuseum and an annotation acquired through Accurator is depicted in Figure 3. Constructs from the Europeana Data Model[5] are used to model the metadata describing the object. An aggregation connects the metadata of the object with a digital representation of the object, in this case, an image. The identifier of the object is connected to metadata such as the title of the object and its subject matter. For this print, the subject matter is an Iconclass concept, representing owls.

Contributors extend the existing information by adding annotations. New annotations are modeled according to the Web Annotation Data model[6], shown as the white ovals in the figure. One annotation has as a target the object, as well as an area of the digital representation of the object. Coordinates formalizing this area correspond to the bounding box drawn by the contributor. The body of the annotation corresponds to the value selected by the contributor, in this case, a concept from the IOC bird list, with the scientific name *Bubo bubo*.

Using concepts instead of plain text to store annotations has a number of advantages. Concepts can have multiple labels, in different languages. The bird on the print of Figure 3 can, for example, be identified by its scientific name *Bubo bubo* as well as its common name in English, Eurasian eagle-owl. When a contributor enters one of these values, they refer to the same species concept. The common name in Dutch can now be used to retrieve the annotated object. This is a significant advantage over annotation using plain text since the annotations do not have to be translated every time a new language is supported. The hierarchy encountered in some vocabularies has additional benefits, for annotation and subsequent object retrieval. More general concepts can be used during annotation at the moment a contributor cannot pinpoint a specific concept. During retrieval, the tree structure can be leveraged in the other direction, if someone searches for a general concept, more specific concepts lower in the hierarchy can be included in the results as well.

Contributors can explore the collection loaded using the semantic search functionality of the Accurator tool. The search is based on a graph search algorithm, which matches keyword queries with labels in the triple store. The graph structure is used to find connected objects and clusters similar objects together (Wielemaker et al., 2008). Users can use this search functionality to explore the collection and find objects to annotate. Search thus functions alongside task assignment as an additional way of accessing tasks. The search algorithm is adapted to interpret added annotations as subject matter metadata, which allows users to directly inspect the result of their efforts. Observing that annotations improve the accessibility of the collection can be an added incentive to keep contributing.

The Accurator annotation tool provides multiple options to export data: annotations can be queried and exported to spreadsheets or RDF files. A public endpoint is available for queries and this can for example be used by systems that integrate multiple cultural heritage collections (Dijkshoorn et al., 2017). Additionally, the annotations are stored using a version management repository which can be easily published online, thereby making the results available to others.

[5]https://pro.europeana.eu/edm-documentation

[6]https://www.w3.org/TR/annotation-model/

Table 1. Overview of the characteristics of the three case studies.

Domain	birds on art	bible prints	fashion images
Goal of campaign	improve access	support comparative research	improve access investigate vocabulary use
Objects to annotate	2,160 artworks *(Rijksmuseum)* 406 prints *(Naturalis)*	246 bible prints *(University Library Vrije Universiteit Amsterdam)*	5,480 fashion objects *(Rijksmuseum)*
Properties to annotate & candidate vocabularies	33,799 taxons *(IOC bird list)* 2 genders 3 stages of life iconography	462 characters *(Bible ontology)* 5,954 themes *(Iconclass)* 34 emotions *(Emotion list)*	717 types *(Fashion thesaurus)* 235 materials *(Fashion thesaurus)* 117 techniques *(Fashion thesaurus)* 20 colors *(Fashion thesaurus)* style period
Niche community	14 bird-watchers	7 bible experts	18 fashion experts
Tool	annotate.accurator.nl	bijbel.accurator.nl	annotate.accurator.nl
Event	birdwatching event *(4-10-2015 Rijksmuseum)*	bible event *(4-4-2016 University Library)*	stitch by stitch event *(23-4-2016 Rijksmuseum)*
Quality assessment	comparison to gold standard	review by professional	sample review by professionals

5. VALIDATION OF NICHESOURCING METHODOLOGY

We validate the Accurator nichesourcing methodology using three real-world case studies in the form of nichesourcing campaigns. These show that the methodology is applicable in the highly different domains of birds on art, bible prints and fashion images. Table 1 provides a schematic overview of the cases, including links to online instances of the tool. In the following subsections, we describe each case in detail and discuss how the Accurator methodology and tool were implemented, listing the individual stages and steps of the methodology. In Section 6, we provide an evaluation of quality and the quantity of the resulting annotations.

5.1. Case study I: Birds on art

The first case study regards birds depicted on objects of the Rijksmuseum Amsterdam[7]. Subject matter is diverse and sometimes outside the area of expertise of the museum's catalogers, who mostly have an art-historical background. At times this results in overly general descriptions, such as the description of the Japanese print of Figure 4: "blue-headed bird, near red vine". In 2015, the museum conducted a nichesourcing campaign, to identify birds on art. Below we show how the four stages of the Accurator methodology are applied in this case study.

Orientation stage In collaboration with the museum, we involved experts in the process of accurately describing subject matter in order to improve access to the collection for online visitors *(determine goal)*. The first type of subject matter that the museum tried to address regarded birds. The query functionality of the museum's collection management system was used to define a set of artworks depicting birds *(identify objects)*. In this case, existing descriptions served as a sufficient basis to identify 2,160 objects. The main goal of the campaign was to accurately identify the depicted species and add this to the objects' metadata *(determine properties and select vocabularies)*. The IOC World Bird List, a comprehensive taxonomy of birds, was identified as candidate vocabulary. Other properties regarded the gender and age of the identified bird. The museum was

[7]https://www.rijksmuseum.nl/en

Figure 4. ***Print by Kono Bairei, titled "Bird and red vine".***

interested whether the contributors could identify iconographic information related to the depicted birds as well.

Many bird-watchers go out into nature every week to seek birds. The museum identified them as the group of enthusiast that it was looking for *(identify niche community)*. To be able to address potential contributors within the niche community, the Naturalis Biodiversity Center[8] was contacted. This natural history museum has access to many communities, including bird-watchers. Naturalis provided an additional set of 406 prints with realistic depictions of birds, which were already annotated by the head of the vertebrate collection and could serve as gold standard for evaluation purposes.

Implementation stage The Accurator annotation tool was deployed on a server and an export of metadata of the set of objects was loaded[9], along with a conversion of the bird list[10] *(deploy tool)*. The tagline and images were changed to suit the bird domain. Short instructions for the annotation fields were written and the bird list was related to the scientific name and common name fields *(deploy tasks)*. A questionnaire inquiring about the experts' experience annotating artworks was created, to be handed out after the annotation session *(setup user feedback elicitation)*. The posed questions regard the usability of the tool, in addition to inquiries about domain-specific adaptions of the tool.

The campaign was marketed as "Birdwatching in the Rijksmuseum" and the event was scheduled to coincide with World Animal Day, making it easier to market *(formulate marketing strategy)*. A page was created on the museum's website[11], advertising the event and annotation tool. The biodiversity center spread the invitation to appropriate channels and the event was picked up by national broadcasters.

[8] https://www.naturalis.nl/en/

[9] https://www.rijksmuseum.nl/en/api/rijksmuseum-oai-api-instructions-for-use

[10] https://github.com/rasvaan/ioc

[11] Website advertising the event: https://www.rijksmuseum.nl/vogelen

Execution stage Two pilot events preceded the event, to test the stability of the system and to make employees of the biodiversity center and museum familiar with the system *(run pilot)*. The two successful pilot events resulted in small incremental updates of the system, after which the organization of the main event could start. The birdwatching event was the first event organized as part of a nichesourcing campaign and set to take place in the historical library of the Rijksmuseum *(organize events)*. To give experts an incentive to join the event, it was accompanied by various presentations related to the subject. After these talks, two and a half hours were spent annotating objects. The annotation session was closed by a curator of the museum after which people could join a bird-oriented guided tour through the museum.

Fourteen bird-watchers annotated objects during the event. Many of them brought their own books of reference (in this case bird guides) and they often formed small groups, among which tasks were discussed. For many, this was a slow paced-process, annotations were thoroughly contemplated and values for all requested data were given when possible. A flyer explained how experts could use the system at home *(run tasks online)*. Unfortunately, as part of the lessons learned, we realized it was a missed opportunity to not have advertised the online system, by sending a follow-up email to report on the results of the event and invite people to continue annotating.

Evaluation stage We used the gold standard of the Naturalis-provided prints to assess annotation quality *(assess quality)*. It was not possible to feed back the results of the campaign into the collection management system of the museum, since adaptions had to be made to allow representing scientific species *(measure impact)*. Comments on the functionality of the annotation system were collected during the event and using the questionnaire. The annotations and questionnaires were analyzed *(analyze user feedback)* and in Section 6 we discuss the results in more detail.

5.2. Case study II: Bible prints

The second case study concerns 18th-century picture bibles. In collaboration with historians and the university library of the Vrije Universiteit Amsterdam[12], a nichesourcing campaign was conducted to enable a comparison of bibles, belonging to the Dutch Protestant heritage collection of the library. Below we describe the four stages of the nichesourcing campaign conducted in 2016.

Orientation stage A peculiar thing about picture bibles is that a buyer could commission which prints should accompany the religious texts (Stronks, 2011). The prints depict bible scenes and were created by renowned artists. Figure 5 shows a digitized bible print from the collection. Analyzing which prints are included can shed light on aspects such as the popularity of artists as well as bible themes *(determine goal)*. For the historians to be able to compare bibles, the pages and the prints among them had to be annotated. The historians selected two bibles that would be interesting to compare: one printed in 1728 by de Hondt and one printed in 1729 by the brothers Keur *(identify objects)*. On request, pages of the two bibles were scanned by a company, resulting in a total of 1,003 images.

The priority of the researchers and the university library was to gather data about the subject matter of prints *(determine properties and select vocabularies)*. Two suitable structured vocabularies were

[12]https://www.ub.vu.nl

Figure 5. Print from Keur bible, depicting multiple biblical themes.

identified for providing auto-completion alternatives: the bible ontology[13] is a source of biblical characters and the Iconclass vocabulary[14] includes descriptions of many biblical themes. The historians were also interested in exploring changes in emotional expressivity depicted on the prints. To allow annotation of emotions, a new vocabulary was created, based on a list of emotions of 18th-century theater texts, composed by the historians[15].

For annotating the subject matter of bible prints, an expert has to be knowledgeable about bible scripture *(identify niche community)*. The collaboration with the university library led to a fitting niche community. The library regularly organizes seminars for "friends of the university library", which often revolved around biblical topics. Since these friends of the library were willing to attend events, the library anticipated that they might also be willing to join annotation events.

Implementation stage The Accurator annotation tool was installed on a university server and customized to accommodate the bible domain. Available metadata was exported from the library catalog[16] and loaded in the annotation tool, together with the three candidate vocabularies *(deploy tool)*. Tasks were defined by adding the fields biblical person, theme and emotion. These fields were related to parts of the candidate vocabularies and a description of the request *(deploy tasks)*. The questionnaire used for the bird domain was adapted, now inquiring about the experience of annotating biblical themes, characters and emotions *(setup user feedback elicitation)*. The library contacted the bible experts and dedicated seminars to annotation events *(formulate marketing strategy)*.

Execution stage A pilot event was organized, during which two talks given by historians provided introductions to crowdsourcing and emotions in picture bibles *(run pilot)*. This did not leave enough

[13]https://bibleontology.com
[14]https://www.iconclass.nl
[15]https://github.com/LaraHack/emotion_ontology
[16]https://github.com/VUAmsterdam-UniversityLibrary/ubvu_bibles

time for an extensive annotation session, although subsequent communications with the participants led to a number of observations. Annotating bible prints required elaborate instructions about the depth and thoroughness of requested annotations. Furthermore, we observed that bible experts are not necessarily experts in recognizing depicted 18th-century emotions. Additionally, many of the digitized pages where either blank or contained only text, making it nonsensical to ask experts to annotate subject matter.

The subset selection and information to gather was adapted based on the pilot event. The task of annotating emotions was removed, to be accomplished at a later time by some other niche community. The digitized pages were classified with whether the page depicts a biblical scene. This was another annotation task but did not require expert knowledge and hence this task was accomplished using a regular crowdsourcing campaign. 246 pages depicted biblical themes and were included for the remainder of the nichesourcing campaign. In addition, a detailed step-by-step instruction manual was created to instruct people on how to use the annotation tool.

For the main annotation event, the introduction was shortened, leaving more room for annotating prints *(organize events)*. A computer room of the university library was used to host the event, with the addition of a hands-on experience with the two original historical picture bibles. Eight friends of the university library attended the annotation event and spent two and a half hours annotating bible prints. After the annotation event, participants were informed about the results of the annotation event and invited to further contribute using the online annotation tool *(run tasks online)*.

Evaluation stage The annotations resulting from the events and from participants continuing at home were reviewed by library staff *(assess quality)*. Verified annotations were exported from the annotation system, published and used by the library *(measure impact)*. The library imported the annotations in its catalog, which now allows browsing based on biblical characters and themes[17]. Input for subsequent events was obtained during the event and from the questionnaires *(analyze user feedback)*.

5.3. Case study III: Fashion images

The third case study regards fashion images. In spring 2016, the Rijksmuseum organized an exhibition called Catwalk, during which fashion objects such as dresses and costumes were displayed[18]. The museum chose this well-advertised exhibition as the context for a nichesourcing campaign.

Orientation stage The goal of the campaign was to better describe fashion objects, thereby improving online access *(determine goal)*. Besides the museum, another user took interest in the annotations. A second goal was to support a researcher who develops a fashion thesaurus and wanted to investigate which terms contributors use to describe fashion objects. The types of objects in the fashion domain are more diverse than the prints and paintings from the previous two case studies. The museum owns a wide range of historical fashion objects, ranging from the dress depicted in Figure 6, to jewelry and prints from fashion magazines. Since the domain included such diverse

[17]This link, for example, lists all prints with a depiction of Moses: https://imagebase.ubvu.vu.nl/cdm/search/collection/bis/searchterm/mozes/

[18]https://www.rijksmuseum.nl/en/catwalk

***Figure 6.** Dress with train, anonymous.*

types of objects, multiple subsets were identified as relevant to the fashion domain, amounting to a total of 5,480 objects *(identify objects)*.

While the objects are diverse, an information specialist of the museum determined that the information that can be gathered about the objects can be categorized under general topics *(determine properties and select vocabularies)*. These topics included technique, material, style period and color. A survey of structured vocabularies resulted in multiple possible candidates per topic, including the Art and Architecture Thesaurus (AAT) of the Getty[19]. For the campaign, it was however decided to use the fashion thesaurus created by Europeana[20], which is based on the AAT, but focusses more specifically on the fashion domain.

The diversity of the fashion domain makes it harder to pinpoint one niche community that is knowledgeable about all facets of the domain *(identify niche community)*. The community of fashion enthusiast (fashionistas) is interested in fashion in a broader sense, but they might not know much about historical objects. There are many experts working with fashion on a professional level, but describing a shoe is something completely different from describing a lace detail. Therefore, to cover as much of the diverse fashion domain as possible, the museum had to turn to a more heterogeneous group of people than the previous two case studies.

Implementation stage The collection data and structured vocabularies were loaded in the Accurator tool the Rijksmuseum already used for the birdwatching event *(deploy tool)*. For the fashion domain, six sub-domains were added: jewelry, accessories, fashion prints, paintings, costumes and lace. Contributors could choose one of these subdomains or the general fashion domain (which includes all objects of the sub-domains), to start adding annotations to. Similar tasks were deployed for each of these sub-domains, relating parts of the Europeana fashion thesaurus to requests for information regarding technique, material, style period and color *(deploy tasks)*.

[19]https://www.getty.edu/research/tools/vocabularies/aat/

[20]https://skos.europeana.eu/api/collections/europeana:fashion.html

Table 2. Results of the three case studies.

Domain	birds on art	bible prints	fashion images
Number of annotations event	835	244	1,357
Number of annotations online	307	2,138	48
Total number of annotations	1,142	2,382	1,405
Quality assessment	comparison to gold standard	review by professional	sample review by professionals
Percentage considered correct	83%	96%	84%

A questionnaire focussed on the fashion domain was created to elicit feedback *(setup user feedback elicitation)*. The event was marketed using the name "Stitch by Stitch"[21] and the organization ModeMuze[22] was willing to help address niche communities *(formulate marketing strategy)*. ModeMuze is a Dutch aggregator of digitized fashion collections. This community was addressed and invited to participate in the event. Additionally, the Catwalk exhibition concluded with a conference for fashion professionals from the cultural heritage sector. An invitation was sent to these professionals as well. The event was organized following the conference, allowing professionals that attended the conference to join the annotation event.

Execution stage Two fashion professionals participated in a small pilot, which did not bring major problems to light *(run pilot)*. The main annotation event took place in the library of the museum *(organize events)*. Since many of the contributors attended a conference in the days preceding the event, introductory talks were limited to an introduction of the annotation tool, leaving plenty of time to annotate objects. The broad invitation to different niche communities led to a diverse group of 18 contributors, including tailors, fashion curators and fashionistas. All of these contributors were asked to join in a discussion at the end of the event, discussing the campaign. The annotation tool stayed online following the event *(run tasks online)*.

Evaluation stage To assess the annotation quality, three fashion professionals evaluated a sample of the collected annotations *(assess quality)*. The free text annotations were compared with the structured vocabulary, seeing whether strong differences occurred, serving as input for the researcher interested in developing a fashion thesaurus for the cultural heritage domain *(measure impact)*. Furthermore, from the discussion following the event and the questionnaires filled in during the event, we received rich feedback on the annotation tool and what information about fashion can be collected *(analyze user feedback)*. The results of the analysis of the questionnaires and the annotations are given in the next section.

6. RESULTS

In this section, we discuss the results of the three nichesourcing campaigns and provide links to the annotation datasets. The quantity and quality of annotations provided by contributors are analyzed in Section 6.1. Section 6.2 comprises the outcomes of a user evaluation of the Accurator annotation tool, which supported the campaigns.

[21] Website advertising the event: https://www.rijksmuseum.nl/stitch-by-stitch

[22] https://www.modemuze.nl

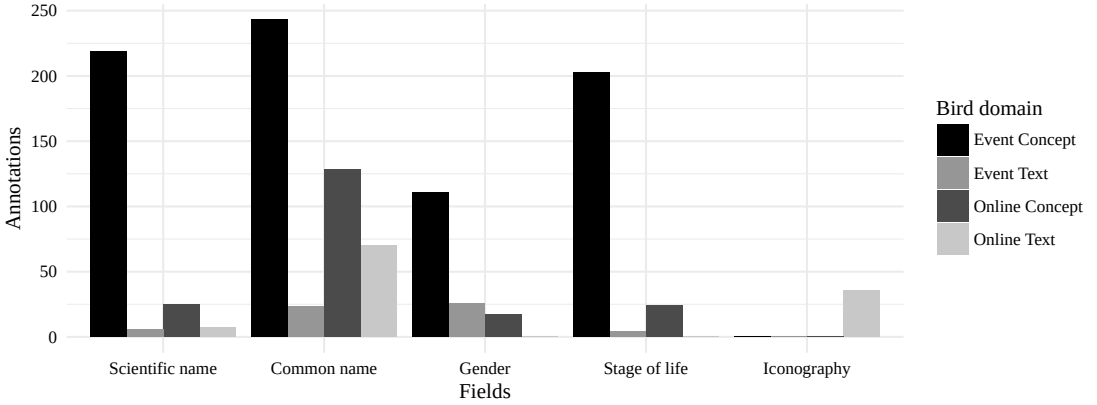


Figure 7. The number of annotations provided by contributors during the birds on art nichesourcing campaign, split by field, type of input and context of data entry.

6.1. **Analysis of the annotations**

For each case study, we analyze the annotations provided in terms of two dimensions: the number of provided annotations and the quality of the annotations. For the quantitative analysis, we split the numbers by annotation field. This provides an indication of whether a field was suitable for a domain. Furthermore, each of these fields is split according to the type of input provided, differentiating between text input and the input of concepts from vocabularies. This shows whether a vocabulary covered the values adequately. The last differentiating factor is the moment the annotation was entered. This is either during an event or during a subsequent possibility to add annotations online. This allows comparing the effectiveness of the campaigns of the three case studies. Regarding the qualitative analysis, for the bird case study, a gold standard was available, allowing validation of the species annotations. The bible annotations and a subset of the fashion annotations were reviewed by professionals, providing an indication of their validity. An overview of the results is given in Table 2.

The birds on art nichesourcing campaign resulted in a total of 1,142 annotations[23], of which 835 annotations were entered during the event and 307 online. The contributors entered on average 59.6 annotations during the event. 65% of the annotations concern species and 85% of these 721 species annotations are concepts from the IOC bird list. During the annotation event, slightly more common names (266) than scientific names (225) were entered. The opposite can be observed of the annotations entered online, here there are 198 common names and 32 scientific names entered. The iconography field is rarely used: During the event nothing was entered in this field, while online the field was used 36 times. The annotations provided during the event mainly concerned the Naturalis collection, containing prints which were not an artistic interpretation of a bird, hence the low count of iconography. A total of 231 stages of life annotations were added and 154 gender annotations. Concepts were used for the vast majority of these annotations.

[23]Repository containing the bird annotations: https://github.com/Rijksmuseum/accurator_annotations

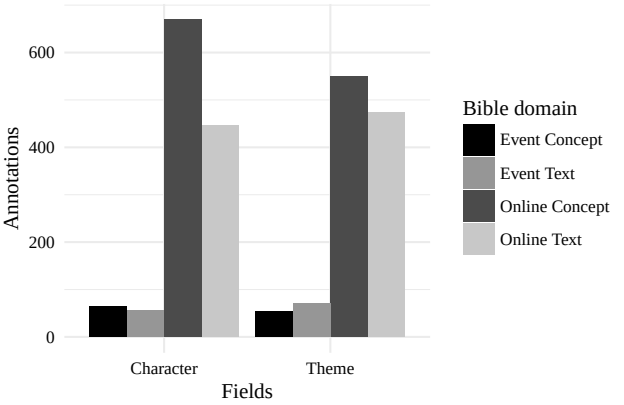


***Figure 8.** The number of annotations provided by contributors during the bible prints nichesourcing campaign, split by field, type of input and context of data entry.*

The Naturalis prints allowed for evaluating the quality of the provided annotations since the depicted species were already annotated by the head of the vertebrate collection. We compare the annotations entered by contributors to this gold standard and distinguish two types of matching. The first type is a direct match of the species concept provided by the professional and the annotation of the contributor. The second type of matching concerns concepts provided by the contributors that are one step higher up in the species taxonomy, thereby matching on a more generic level with the depicted species. Out of the 427 species annotations added to a print with gold standard, 344 of the annotations (80%) exactly match with the annotation of the professional and 11 annotations (3%) match on a more general level. Most errors regard the misidentification of species within the same taxonomic family as the correct species. Examples are the identification of a hen harrier instead of a pallid harrier and a ring-necked duck instead of a tufted duck. The high number of correct annotations by a niche community is in-line with results observed in online groups determining species of sea slugs (Chamberlain, 2014).

2,382 annotations were obtained during the bible prints campaign[24]. An overview of the obtained annotations is given in Figure 8. The event resulted in 244 annotations, a contributor added 34.9 annotations on average. In contrast to the other two domains, which have a low number of annotations added online, 90% of the bible annotations were obtained online. In total, 1,236 biblical characters were annotated, slightly more than the 1,146 themes. Vocabulary concepts were more often used than text annotations: 56% of the annotations. However, the use of concepts from structured vocabularies is lower than for example the species annotations within the bird domain.

In July 2016, personnel of the university library reviewed all annotations available at that moment: 1,455 annotations in total. 96% of the annotations were accepted: 630 theme and 764 bible character annotations. Errors were of a different nature than those of the birds on art campaign. Some themes and characters were misidentified, but most incorrect annotations were a result of annotators adding a dash in a field when they did not know which annotation to provide. The increase

[24] Annotation repository: https://github.com/VUAmsterdam-UniversityLibrary/ubvu_bible_annotations

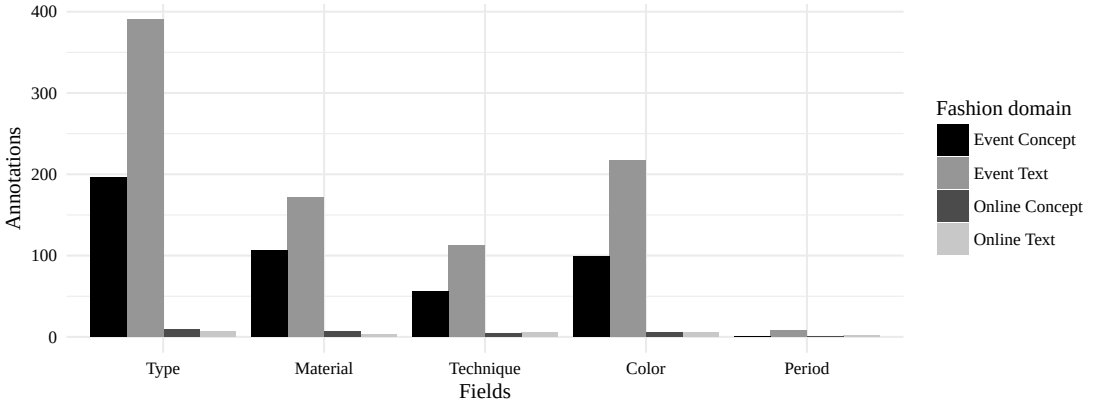


Figure 9. The number of annotations provided by contributors during the fashion images nichesourcing campaign, split by field, type of input and context of data entry.

of text annotations also resulted in more rejected annotations due to spelling errors. The verified annotations have been added to the libraries' catalog.

A total of 1,405 annotations were added during the fashion images campaign[25], as shown in Figure 9. Just 48 annotations were a result of the online campaign, 97% of all annotations were a result of the event. During this event, contributors added 75 annotations on average. 602 annotations concerned types of objects, 288 materials, 179 techniques and 326 colors. Style periods were rarely added, just 10 times. In contrast to the other two domains, the use of concepts is low: 34% of the total annotations originate out of the Europeana Fashion Thesaurus, the rest are textual annotations.

A sample of 40 annotations was evaluated by 3 fashion professionals. One professional works at the Rijksmuseum, one works for the fashion aggregator Modemuze and the last for a fashion museum in Antwerpen. Ten annotations were randomly picked from respectively the type, material, technique and color annotations. For each annotation, the professionals were asked whether it was correct, incorrect, or whether they were unable to assess it. We used majority voting to reach an assessment for 37 of the annotations, for 3 annotations the evaluations were inconclusive. Out of the sample, 89% of type annotations, 78% of the material annotation, 78% of the technique annotations and 90% of the color annotations were judged to be correct. From the in total 37 annotations upon which agreement was reached, 84% were considered correct. Incorrect annotations regarded misidentified concepts (e.g. an annotator added the material satin instead of silk), annotations added to the wrong annotations field (e.g. an annotator added the material leather to the technique field) and spelling mistakes.

6.2. Evaluation Accurator annotation tool

In order to evaluate the effectiveness of the Accurator annotation tool, at the end of each of the annotation events, questionnaires were handed out. 14 birdwatchers, 9 bible experts and 18 fashion

[25]Repository containing the fashion annotations: https://github.com/Rijksmuseum/accurator_annotations

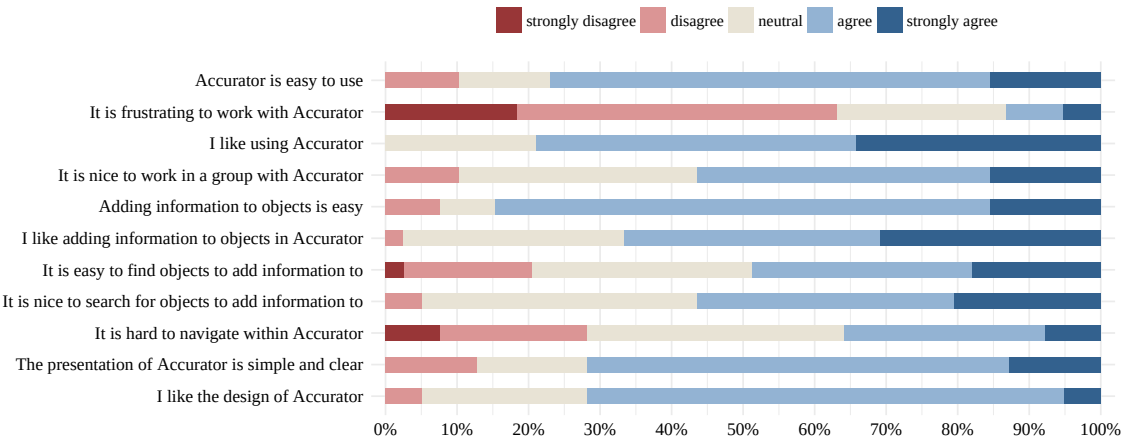


Figure 10. ***Overview of the answers provided that regard the usability of the Accurator annotation tool.***

experts filled in the questionnaire. In this section, we list the outcomes, focussing on the discussion around task assignment and the usability of the Accurator annotation tool.

During the three campaigns, different settings for task assignment were used, which are described in Section 4.2. The bible prints domain used the ranked setting, the fashion images domain the sub-domain based setting and the birds on art domain used recommendation. Since the latter two are more advanced ways of assigning tasks to contributors, questions of how these settings were experienced by the contributors were included. The sub-domain based setting of the fashion images domain is deemed useful by 89% of the respondents. Contributors comment that using the sub-domains it is easier to access objects they know something about. Some would like the option to refine a sub-domain by adding filters, thereby, for example, indicating the type of accessories recommended. Additionally, dividing the domains based on style period would be appreciated. 79% of the bird watchers find recommendation based on expertise useful. They comment that it makes the process more efficient, although a different elicitation of expertise is proposed by many contributors. Expertise topics concerned different families of the biological taxonomy, while many contributors think it would be more useful to ask how much someone knows about a certain region where the birds reside.

The questionnaire included 11 statements regarding the usability of the annotation tool. Participants were asked to indicate their agreement on a five-level Likert scale, ranging from “strongly disagree” to “strongly agree”. Figure 10 shows an overview of the answers of the contributors of the three case studies combined, which sums up to a total of 41 questionnaires. The evaluation of the general usability of the tool is good, 77% of the contributors agree that Accurator is easy to use. 63% disagrees with that it is frustrating to use the tool. Comments of contributors that found the tool frustrating to use regarded problems with the interface, the inability to closer inspect objects using different angles and the challenge of finding objects to annotate. A high number of contributors (79%) did like to use Accurator, no one disagreed on this point. Over half of the respondents

enjoyed working with the tool in a group, although an additional one-third of the responses regards this point as neutral.

The questionnaires also included more focused questions, regarding the ability to add information to objects, searching for objects and the design of the tool. 85% of the respondents find it easy to add information to objects, two-thirds of the respondents like adding information to objects using the tool. Navigating to tasks is deemed more complicated, over half of the respondents disagreed or replied neutrally on the question whether it was easy to find an object to add information to. Over half of the contributors (56%) did enjoy searching for objects. Navigating the tool is found to be more complicated, 36% agreed that it is hard to navigate Accurator. Over two-thirds of the contributors like the design of Accurator and find the tool simple and clear.

7. DISCUSSION & FUTURE WORK

Nichesourcing is a method for outsourcing tasks that require a significant level of expertise in a specific domain. The Accurator nichesourcing methodology presented in this paper is geared towards executing knowledge-intensive annotation tasks in the cultural heritage domain in a sustainable and repeatable fashion. The involvement of niche communities with a specific domain of interest and off-line annotation events are central to the Accurator methodology. The Accurator annotation tool supports the methodology and a user evaluation indicates that the design and usability of the tool are appreciated, as well as working together with other members of the community. The three case studies show that the nichesourcing methodology in combination with the annotation tool can be used to collect high-quality annotations in a variety of domains.

While all three case studies required experts to be knowledgeable about the domain on hand, annotating fashion images proved to be the most challenging. Determining materials and techniques from single images is difficult and the formulated requests for annotations proved to be more ambiguous as well. The abstract nature and imagined aspects of some of the images in the birds on art domain, made pinpointing the exact species challenging at times. Furthermore, the use of terms from structured vocabularies differs significantly per case study. The number of concepts used is an indicator of how suitable a vocabulary is to describe a property of a collection object. The difference in collected annotations between the event and the online tool underlines the importance of a strong marketing strategy. After the bible annotation event, multiple emails were sent inviting people to keep contributing, which clearly shows in the results.

All case studies presented in this paper are conducted in the cultural heritage domain. Although this focus prohibits us from making definitive claims about the generalizability of our findings, we expect that the methodology and tool can be utilized in cases from other domains as well. The applicability and usefulness will depend on the characteristics of the case in question. Relevant characteristics concern the dataset, vocabularies and the community. We will briefly discuss each of these below.

At the foundation of the Accurator nichesourcing methodology lies the availability of a dataset with metadata about objects. The metadata can describe real-world objects or born-digital objects, although digital representations of the objects should be available to be annotated. Examples of other usable datasets are a product catalog of an online shop or sound clips from a radio broadcaster. In this research, we dealt with a single modality of representations: images. We expect that the

Accurator nichesourcing methodology can also be used for other media modalities, even though the Accurator annotation tool will have to be adapted. A broadcaster might need to, for example, extend the tool with functionality to support annotating sound or video clips.

Furthermore, the degree in which structured vocabularies cover the domain influences the accuracy of enrichments. As discussed in this paper, there is a number of high-quality vocabularies available in the cultural heritage domain, which is not the case for every domain. For the domains that lack appropriate vocabularies, more generic external datasets could be used, such as WikiData (Vrandečić and Krötzsch, 2014). An active community interested in the topic at hand should exist for the nichesourcing methodology to be effective since it relies on the voluntary contribution of niche groups. Not every case will have a community cut out for it, although we recommend requesters to be inventive in the selection of communities. A broadcaster looking to describe video clips might, for example, involve locals knowledgeable about the situation.

In future campaigns, we plan to optimize settings that impact the results, such as the number of selected objects, the formulation of information requests and the influence of the marketing schedule. Additionally, we will further analyze the impact of different types of task assignment. By comparing different approaches within the same domain, we can measure the effects of task assignment on dimensions such as accuracy and time spent per provided annotation. Furthermore, we plan to translate the social aspect of the annotation events into the functionality of the tool and investigate whether this will retain more contributors.

To accomplish goals more efficiently, we will investigate embedding nichesourcing in hybrid crowdsourcing workflows, splitting a campaign into subtasks that are solved using different methods within the human computation spectrum. This would have the benefit that for simple tasks, that can be solved by anyone in the crowd, we can resort to methods other than nichesourcing, thereby not wasting the goodwill of our expert volunteers. Other possibilities are automating parts of the campaign, such as utilizing computer vision to recognize objects on images. Finding a hybrid approach that strikes the right balance of quality and quantity of annotations will improve the usefulness of cultural heritage data published online.

Acknowledgements This publication was supported by the Dutch national program COMMIT/. We thank all members of the SEALINCMedia project for their input.

8. REFERENCES

Albert, W and Tullis, T. (2013). *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics* (2nd ed.). Morgan Kaufmann.

Bevan, A, Pett, D, Bonacchi, C, Keinan-Schoonbaert, A, Lombraña González, D, Sparks, R, Wexler, J, and Wilkin, N. (2014). Citizen Archaeologists. Online Collaborative Research about the Human Past. *Human Computation Journal* 1, 2 (2014), 185–199. DOI: http://dx.doi.org/10.15346/hc.v1i2.9

Ceolin, D, Nottamkandath, A, and Fokkink, W. (2012). Automated evaluation of annotators for museum collections using subjective logic. In *Proceedings of the 6th IFIP Trust Management Conference (IFIPTM '12)*. Springer, Berlin, Heidelberg, 232–239. DOI: http://dx.doi.org/10.1007/978-3-642-29852-3_18

Chamberlain, J. (2014). Groupsourcing: distributed problem solving using social networks. In *Proceedings of the 2nd AAAI Conference on Human Computation and Crowdsourcing (HCOMP '14)*. The AAAI Press, Palo Alto, CA, USA, 22–29.

Chun, S, Cherry, R, Hiwiller, D, Trant, J, and Wyman, B. (2006). Steve.museum: an ongoing experiment in social tagging, folksonomy, and museums. In *Proceedings of the Museums and the Web conference*, Jennifer Trant and David Bearman (Eds.). http://www.archimuse.com/mw2006/papers/wyman/wyman.html

Cosley, D, Frankowski, D, Terveen, L, and Riedl, J. (2007). SuggestBot: using intelligent task routing to help people find work in Wikipedia. In *Proceedings of the 12th International Conference on Intelligent User Interfaces (IUI '07)*. ACM, New York, New York, USA, 32–41. DOI:http://dx.doi.org/10.1145/1216295.1216309

de Boer, V, Hildebrand, M, Aroyo, L, Leenheer, P. D, Dijkshoorn, C, Tesfa, B, and Schreiber, G. (2012)a. Nichesourcing: harnessing the power of crowds of experts. In *Proceedings of the 18th International Conference on Knowledge Engineering and Knowledge Management (EKAW '12)*, Annette ten Teije, Johanna Völker, Siegfried Handschuh, Heiner Stuckenschmidt, Mathieu d'Acquin, Andriy Nikolov, Nathalie Aussenac-Gilles, and Nathalie Hernandez (Eds.). Springer, Berlin, Heidelberg, 16–20. DOI:http://dx.doi.org/10.1007/978-3-642-33876-2_3

de Boer, V, Wielemaker, J, van Gent, J, Hildebrand, M, Isaac, A, van Ossenbruggen, J, and Schreiber, G. (2012)b. Supporting Linked Data Production for Cultural Heritage Institutes: The Amsterdam Museum Case Study. In *Proceedings of the 9th Extended Semantic Web Conference*, Elena Simperl, Philipp Cimiano, Axel Polleres, Oscar Corcho, and Valentina Presutti (Eds.). ESWC '12, Vol. 7295. Springer Berlin Heidelberg, Berlin, Heidelberg, 733–747. DOI:http://dx.doi.org/10.1007/978-3-642-30284-8_56

Difallah, D. E, Demartini, G, and Cudré-Mauroux, P. (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:http://dx.doi.org/10.1145/2488388.2488421

Dijkshoorn, C, Bucur, C.-L, Brinkerink, M, Pieterse, S, and Aroyo, L. (2017). DigiBird: On The Fly Collection Integration Supported By The Crowd. In *Proceedings of the Museums and the Web conference*. http://mw17.mwconf.org/paper/digibird-on-the-fly-collection-integration-supported-by-the-crowd/

Dijkshoorn, C, Leyssen, M. H. R, Nottamkandath, A, Oosterman, J, Traub, M, Aroyo, L, Bozzon, A, Fokkink, W, Houben, G.-J, Hovelmann, H, Jongma, L, van Ossenbruggen, J, Schreiber, G, and Wielemaker, J. (2013). Personalized Nichesourcing: Acquisition of Qualitative Annotations from Niche Communities. In *Workshop Proceedings of the 21st Conference on User Modeling, Adaptation, and Personalization (CEUR Workshop Proceedings)*, Shlomo Berkovsky, Eelco Herder, Pasquale Lops, and Olga C. Santos (Eds.), Vol. 997. CEUR-WS.org.

Doan, A, Ramakrishnan, R, and Halevy, A. Y. (2011). Crowdsourcing systems on the world-wide web. *Commun. ACM* 54, 4 (April 2011), 86–96. DOI:http://dx.doi.org/10.1145/1924421.1924442

Ellis, A, Gluckman, D, Cooper, A, and Andrew, G. (2012). Your Paintings: a nation's oil paintings go online, tagged by the public. In *Proceedings of the Museums and the Web conference*. http://www.museumsandtheweb.com/mw2012/papers/your_paintings_a_nation_s_oil_paintings_go_onl

Gligorov, R, Hildebrand, M, Ossenbruggen, J, Aroyo, L, and Schreiber, G. (2013). An evaluation of labelling-game data for video retrieval. In *Proceedings of the 35th European Conference on IR Research (ECIR '13)*, Pavel Serdyukov, Pavel Braslavski, SergeiO. Kuznetsov, Jaap Kamps, Stefan Rüger, Eugene Agichtein, Ilya Segalovich, and Emine Yilmaz (Eds.). Springer, Berlin Heidelberg, 50–61. DOI:http://dx.doi.org/10.1007/978-3-642-36973-5_5

Gligorov, R, Hildebrand, M, van Ossenbruggen, J, Schreiber, G, and Aroyo, L. (2011). On the role of user-generated metadata in audio visual collections. In *Proceedings of the 6th international Conference on Knowledge Capture (K-CAP '11)*. ACM, New York, New York, USA, 145–152. DOI:http://dx.doi.org/10.1145/1999676.1999702

Goto, S, Ishida, T, and Lin, D. (2016). Understanding crowdsourcing workflow: modeling and optimizing iterative and parallel processes. In *Proceedings of the 4th AAAI Conference on Human Computation and Crowdsourcing (HCOMP '16)*. The AAAI Press, Palo Alto, CA, USA, 52–58.

Inel, O, Khamkham, K, Cristea, T, Dumitrache, A, Rutjes, A, van der Ploeg, J, Romaszko, L, Aroyo, L, and Sips, R.-J. (2014). CrowdTruth: machine-human computation framework for harnessing disagreement in gathering annotated data. In *Proceedings of the 13th International Semantic Web Conference (ISWC '14)*, Peter Mika, Tania Tudorache, Abraham Bernstein, Chris Welty, Craig Knoblock, Denny Vrandečić, Paul Groth, Natasha Noy, Krzysztof Janowicz, and Carole Goble (Eds.). Springer, Cham, 486–504. DOI:http://dx.doi.org/10.1007/978-3-319-11915-1_31

Ipeirotis, P. G and Gabrilovich, E. (2014). Quizz: Targeted Crowdsourcing with a Billion (Potential) Users. In *Proceedings of the 23rd International Conference on World Wide Web (WWW '14)*. ACM, New York, NY, USA, 143–154. DOI:http://dx.doi.org/10.1145/2566486.2567988

Kulkarni, A, Narula, P, Rolnitzky, D, and Kontny, N. (2014). Wish: amplifying creative ability with expert crowds. In *Proceedings of the 2nd AAAI Conference on Human Computation and Crowdsourcing (HCOMP '14)*. The AAAI Press, Palo Alto, CA, USA, 112–120.

Moon, J. Y and Sproull, L. S. (2008). The role of feedback in managing the Internet-based volunteer work force. *Information Systems Research* 19, 4 (2008), 494–515.

Mouromtsev, D, Haase, P, Cherny, E, Pavlov, D, Andreev, A, and Spiridonova, A. (2015). Towards the Russian Linked Culture Cloud: Data Enrichment and Publishing. In *Proceedings of the 12th Extended Semantic Web Conference (ESWC '15)*, Fabien Gandon, Marta Sabou, Harald Sack, Claudia d'Amato, Philippe Cudré-Mauroux, and Antoine Zimmermann (Eds.). Springer International Publishing, Cham, 637–651. DOI:http://dx.doi.org/10.1007/978-3-319-18818-8_39

Noordegraaf, J, Bartholomew, A, and Eveleigh, A. (2014). Modeling crowdsourcing for cultural heritage. In *Proceedings of the Museums and the Web conference*. http://mw2014.museumsandtheweb.com/paper/modeling-crowdsourcing-for-cultural-heritage/

Oomen, J and Aroyo, L. (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, New York, USA, 138–149. DOI: http://dx.doi.org/10.1145/2103354.2103373

Oosterman, J and Houben, G.-J. (2016). On the Invitation of Expert Contributors from Online Communities for Knowledge Crowdsourcing Tasks. In *Proceedings of the 16th International Conference on Web Engineering (ICWE16)*, Alessandro Bozzon, Philippe Cudre-Maroux, and Cesare Pautasso (Eds.). Springer, Cham, 413–421. DOI:http://dx.doi.org/10.1007/978-3-319-38791-8_27

Palmatier, R and Sridhar, S. (2017). *Marketing strategy: Based on first principles and data analytics*. Red Globe Press.

Quinn, A. J and Bederson, B. B. (2011). Human computation: a survey and taxonomy of a growing field. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, New York, USA, 1403–1412. DOI: http://dx.doi.org/10.1145/1978942.1979148

Raddick, M. J, Bracey, G, Gay, P. L, Lintott, C. J, Murray, P, Schawinski, K, Szalay, A. S, and Vandenberg, J. (2010). Galaxy Zoo: exploring the motivations of citizen science volunteers. *Astronomy Education Review* 9, 1 (2010), 1–18. DOI:http://dx.doi.org/10.3847/AER2009036

Ridge, M. (2013). From tagging to theorizing: deepening engagement with cultural heritage through crowdsourcing. *Curator: The Museum Journal* 56, 4 (2013), 435–450. DOI:http://dx.doi.org/10.1111/cura.12046

Sarasua, C, Simperl, E, Noy, N, Bernstein, A, and Leimeister, J. M. (2015). Crowdsourcing and the Semantic Web: a research manifesto. *Human Computation Journal* 2, 1 (2015), 3–17. DOI:http://dx.doi.org/10.15346/hc.v2i1.2

Shadbolt, N, Berners-Lee, T, and Hall, W. (2006). The Semantic Web Revisited. *IEEE Intelligent Systems* 21, 3 (May 2006), 96–101. DOI:http://dx.doi.org/10.1109/MIS.2006.62

Simon, R, Haslhofer, B, Robitza, W, and Roochi, E. M. (2011). Semantically augmented annotations in digitized map collections. In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries (JCDL '11)*. ACM, New York, New York, USA, 199–202. DOI:http://dx.doi.org/10.1145/1998076.1998114

Stronks, E. (2011). *Negotiating differences: word, image and religion in the Dutch Republic*. Vol. 155. Brill.

Szekely, P, Knoblock, C. A, Yang, F, Zhu, X, Fink, E. E, Allen, R, and Goodlander, G. (2013). Connecting the Smithsonian American Art Museum to the Linked Data Cloud. In *Proceedings of the 10th Extended Semantic Web Conference (ESWC '13)*, Philipp Cimiano, Oscar Corcho, Valentina Presutti, Laura Hollink, and Sebastian Rudolph (Eds.), Vol. 7882. Springer Berlin Heidelberg, Berlin, Heidelberg, 593–607. DOI:http://dx.doi.org/10.1007/978-3-642-38288-8_40

Traub, M. C, van Ossenbruggen, J, He, J, and Hardman, L. (2014). Measuring the effectiveness of gamesourcing expert oil painting annotations. In *Proceedings of the 36th European Conference on IR Research (ECIR '14)*, Maarten de Rijke, Tom Kenter, Arjen P. de Vries, ChengXiang Zhai, Franciska de Jong, Kira Radinsky, and Katja Hofmann (Eds.). Springer, Cham, 112–123. DOI:http://dx.doi.org/10.1007/978-3-319-06028-6_10

von Ahn, L and Dabbish, L. (2004). Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04)*. ACM, New York, NY, USA, 319–326. DOI:http://dx.doi.org/10.1145/985692.985733

Vrandečić, D and Krötzsch, M. (2014). Wikidata: A Free Collaborative Knowledgebase. *Commun. ACM* 57, 10 (09 2014), 78–85. DOI: http://dx.doi.org/10.1145/2629489

Wenger, E, McDermott, R. A, and Snyder, W. (2002). *Cultivating communities of practice: A guide to managing knowledge*. Harvard Business Press.

Whitehill, J, fan Wu, T, Bergsma, J, Movellan, J. R, and Ruvolo, P. L. (2009). Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. In *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta (Eds.). Curran Associates, 2035–2043.

Wielemaker, J, Beek, W, Hildebrand, M, and van Ossenbruggen, J. (2016). ClioPatria: a SWI-Prolog infrastructure for the Semantic Web. *Semantic Web Journal* 7, 5 (2016), 529–541. DOI:http://dx.doi.org/10.3233/SW-150191

Wielemaker, J, Hildebrand, M, and Van Ossenbruggen, J. (2007). Using Prolog as the fundament for applications on the semantic web. In *Proceedings of the 2nd Workshop on Applications of Logic Programming and to the web, Semantic Web and Semantic Web Services (CEUR Workshop Proceedings)*, S Heymans, A Polleres, E Ruckhaus, D Pearse, and G Gupta (Eds.). CEUR-WS.org, 84–98.

Wielemaker, J, Hildebrand, M, van Ossenbruggen, J, and Schreiber, G. (2008). Thesaurus-Based Search in Large Heterogeneous Collections. In *Proceedings of the 7th International Semantic Web Conference (ISWC '08)*, Amit Sheth, Steffen Staab, Mike Dean, Massimo Paolucci, Diana Maynard, Timothy Finin, and Krishnaprasad Thirunarayan (Eds.). Springer, Berlin, Heidelberg, 695–708. DOI:http://dx.doi.org/10.1007/978-3-540-88564-1_44

Yadav, P and Darlington, J. (2016). Design guidelines for the user-centred collaborative citizen science platforms. *Human Computation*

Journal 3, 1 (2016), 205–211. DOI:http://dx.doi.org/10.15346/hc.v3i1.11