Summary

Semantic Web technologies allow cultural heritage institutions to publish interconnected, interoperable data, with explicit semantics. The source of the data published by museums is often the basic metadata recorded in systems aimed at collection management. As a consequence, users are deprived of the curated contextual information of regular exhibitions. To address this problem and provide better access to online collections, institutions employ various approaches to improve object descriptions. Among them is crowdsourcing, a quick and inexpensive source of large quantities of descriptions. However, it remains a challenge to ensure the quality of crowdsourced information, especially for knowledge-intensive tasks. In this thesis, we introduce *nichesourcing*, a method to solve knowledge-intensive tasks, by identifying and engaging small groups of experts. We present a five-step method, to enrich and contextualize object metadata using nichesourcing, thereby improving access to online cultural heritage collections.

Analysis of collection data

The first step of the method concerns the analysis of collection data. During this step, we assess the suitability of the chosen data model and the number of references to external datasets. In Chapter 2, we analyze the Linked Data of the Rijksmuseum Amsterdam, as a case study. The Rijksmuseum collection comprises over a million objects, of which only a fraction can be on display at a given time. To open up the remaining collection, the museum started to digitize objects and publish the resulting information online. The Linked Data of the museum consists of over 22 million statements, describing over 350,000 objects, of which more than 207,000 include a reference to an image. The data is used to support search, recommendation, collection integration and browsing.

The Rijksmuseum uses contextual concepts from structured vocabularies to describe objects. While the museum maintains its own vocabularies to preserve its own perspective, an increasing number of contextual concepts is related to external datasets. The collection data is structured using the Europeana Data Model. Not all aspects of the collection can be captured with the modeling constructs recommended by Europeana. Therefore, we discuss modeling challenges and proposed solutions for contextualizing cultural heritage data in the next chapter.

Contextualization of cultural heritage data

Ontologies make the semantics of data explicit, by providing a shared conceptualization. When a cultural heritage institution wants to publish Linked Data, it is confronted with the choice of which ontology to use. This decision has implications for the source data that can be included, as well as the structure of the resulting Linked Data. As part of the five-step method, we focus in Chapter 3 on how ontologies can be used to structure and represent contextual information about objects in cultural heritage collections. We discuss modeling challenges that regard specialization, object- and event-centric approaches, temporality, representations, views and subject matter. For each challenge, we show modeling approaches of two ontologies often used in the cultural heritage domain: the Europeana Data Model and the CIDOC Conceptual Reference Model.

Based on the discussed modeling challenges, we formulate six requirements for cultural heritage ontologies: 1) the ability to specialize an ontology without decreasing its interoperability, 2) support for recording both attributes as well as events related to objects, 3) ability to capture changes over time, 4) ability to separate descriptions of objects and their representations, 5) support for capturing multiple sources describing the same object and 6) possibility to contextualize objects using subject matter. By considering these requirements, institutions can make a more informed choice when deciding on which ontology to use to contextualize data published online.

Nichesourcing

The usefulness of cultural heritage data hinges on the quality and diversity of descriptions of collection objects. In many cases, existing descriptions are insufficient for retrieval and research tasks, resulting in the need for additional 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 challenging for more complex and domain-specific tasks. Nichesourcing addresses this problem, by tapping into the expert knowledge available in niche communities.

In Chapter 4, we present 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 contributions are the following: 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 case studies concern birds on art, bible prints and fashion images. We compare the quality and quantity of obtained annotations, 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.

Diversification of search results

In Chapter 5, we consider whether, and to what extent, additional semantics in the form of Linked Data can support explorative search. As a case study, we use the Linked Data of the Rijksmuseum, extended with various structured vocabularies. We apply an existing graph search algorithm to this data, which finds paths in the graph from the search term to target objects. Next, the algorithm clusters results with similar paths together. We use the number of resulting clusters and the path length as indicators of diversity. As sample queries, we collected the terms in the museum's query log for the duration of one month.

The results show that for this application domain, the added semantics lead to 1) an increase in the number of results, and 2) an increase in the variety of search results. We hypothesize that the following two factors impact the usefulness of vocabularies for search: 1) the number of links between distinct concepts and the metadata objects and 2) the richness of the internal links between concepts in vocabularies. This fourth step of the method illustrates that additional semantics provided by structured vocabularies can help users to explore collections and reach more objects related to their interest.

Integration of collections

Online cultural heritage collections often contain complementary objects, which makes integration of heterogeneous collections a worthwhile effort. In Chapter 6, we present the DigiBird system that provides access to four distinct nature-related collections and reinforces crowdsourcing initiatives. The system is designed to harmonize complementary collection objects, make crowd contributions instantaneously available and allow the monitoring of multiple crowdsourcing systems using one dashboard.

Harmonizing data from multiple systems that adhere to different standards proved to be a challenge. The data originates from dynamic systems, as a continuous stream of crowd contributions alters and extends the datasets. With the DigiBird system, institutions can decide to use data in an early stage of the crowdsourcing process. Additionally, undertaking crowdsourcing projects together allows the sharing of resources and provides insights into the time needed to collect results. By leveraging standardized data models and annotations from structured vocabularies, the DigiBird system illustrates the added value of enrichments and the benefits of Linked Data for collection integration.

Conclusion

In this thesis, we present a method to contextualize and enrich cultural heritage collections, to support explorative search and collection integration. Museums with similar collections as the Rijksmuseum will be able to use the method without requiring major changes. We expect that many steps of the method can be utilized in other domains as well. Disseminating high-quality information about objects is embedded in the mission of cultural heritage institutions. Institutions that want to publish rich contextualized data online, have to assess the quality of external structured vocabularies and find efficient approaches to relate collection data to these new sources. Nichesourcing is one of these approaches, additionally providing ways to engage with the public. To keep contributors motivated, it is essential that they know their work matters. We show the direct impact of annotations on search functionality and collection integration, highlighting the potential of crowd contributions and Linked Data.

Samenvatting

Semantic Web technologie stelt cultureel-erfgoedinstellingen in staat om data te publiceren, waarvan de betekenis is vastgelegd en welke van context is voorzien door middel van externe bronnen. Beperkte metadata over objecten, bedoeld voor het beheren van collecties, vormt veelal de basis voor de gepubliceerde data. De uitgebreide informatie waar tentoonstellingen normaal gezegd in voorzien, ontbreekt daardoor voor gebruikers. Om dit probleem aan te pakken en de toegang tot online collecties te verbeteren, zijn instellingen begonnen met het verbeteren van objectbeschrijvingen. *Crowdsourcing* is een manier om snel grote hoeveelheden beschrijvingen te verzamelen. Het blijft echter een uitdaging om de kwaliteit te borgen van informatie die door crowdsourcing verkregen is. Dit geldt in het bijzonder voor annotatietaken die specifieke kennis vereisen. In deze dissertatie introduceren we *nichesourcing*, een methode voor het uitvoeren van kennisintensieve taken, waarbij gespecialiseerde groepen in het annotatieproces worden betrokken. Deze dissertatie beschrijft een methode om met behulp van nichesourcing objectbeschrijvingen te verrijken en te contextualiseren, waardoor online collecties beter toegankelijk worden. Deze methode bestaat uit vijf stappen.

Analyse van collectiedata

De eerste stap van de methode betreft de analyse van collectiedata. We kijken tijdens deze stap naar de geschiktheid van het gekozen datamodel en het aantal verwijzingen naar externe datasets. In hoofdstuk 2 analyseren we als casestudy de *Linked Data* van het Rijksmuseum Amsterdam. De Rijksmuseum collectie bestaat uit meer dan een miljoen objecten, waar op enig moment enkel een fractie kan worden tentoongesteld. Om ook de rest van de collectie toegankelijk te maken, is het museum gestart met het digitaliseren van objecten en het online publiceren van informatie. De Linked Data van het museum bestaat uit 2.846.996 *statements*, die 351.814 objecten beschrijven, waarvan 207.441 een corresponderende afbeelding hebben. De data wordt gebruikt om de collectie toegankelijk en doorzoekbaar te maken, relevante objecten aan te raden en de collectie te integreren met andere collecties.

Het Rijksmuseum gebruikt concepten van gestructureerde vocabulaires om objecten te beschrijven. Ondanks dat het museum er ook voor kiest een eigen vocabulaire te onderhouden, wordt een toenemend aantal concepten gerelateerd aan externe datasets. De collectiedata wordt gestructureerd met behulp van het Europeana Data Model. Niet alle aspecten van de objecten kunnen echter adequaat worden beschreven met de door Europeana voorgeschreven elementen. Daarom bespreken we de uitdagingen van het modelleren en contextualiseren van cultureel-erfgoeddata in het volgende hoofdstuk.

Cultureel-erfgoeddata contextualiseren

Een ontologie maakt de betekenis van data expliciet, door middel van een gedeelde conceptualisatie. Wanneer een cultureel-erfgoedinstelling Linked Data wil publiceren, moet er worden gekozen welke ontologie het meest geschikt is. Deze beslissing heeft gevolgen voor welke data er kan worden opgenomen in de dataset, maar ook voor de structuur van de resulterende Linked Data. In hoofdstuk 3 kijken we hoe ontologiën gebruikt kunnen worden voor het structureren en representeren van contextuele informatie over objecten in cultureel-erfgoedcollecties. We bespreken uitdagingen op het gebied van data modelleren, met betrekking tot specialisatie, object- of gebeurtenisgerichte aanpakken, tijd, representatie, perspectieven en onderwerpontsluiting. Elke uitdaging illustreren we met de benaderingen van twee veel gebruikte ontologiën in het cultureel-erfgoeddomein: het Europeana Data Model en het CIDOC Conceptual Reference Model.

Gebaseerd op bovenstaande uitdagingen, formuleren we zes vereisten voor cultureel-erfgoedontologiën: 1) een ontologie kan worden gespecialiseerd zonder dat de interoperabiliteit daar onder lijdt, 2) zowel eigenschappen van objecten, als wel gebeurtenissen gerelateerd aan objecten kunnen worden vastgelegd, 3) veranderingen als gevolg van het verstrijken van tijd kunnen worden beschreven, 4) er kan onderscheid worden gemaakt tussen objecten en hun representaties, 5) verschillende bronnen over hetzelfde object kunnen worden vastgelegd en 6) objecten kunnen van context worden voorzien door middel van onderwerpontsluiting. Instellingen kunnen een afgewogen keuze maken over welke ontologie te gebruiken, wanneer ze deze vereisten in overweging nemen.

Nichesourcing

De waarde van cultureel-erfgoeddata hangt af van de kwaliteit en diversiteit van de beschrijvingen van collectieobjecten. In veel gevallen zijn al bestaande beschrijvingen niet geschikt voor het ondersteunen van onderzoek of voor het online toegankelijk maken van de collectie, waardoor extra annotaties nodig zijn. Het uitvoeren van deze annotatietaken wordt bemoeilijkt doordat er vaak domein-specifieke kennis voor nodig is. Waar crowdsourcing vaak succesvol kan worden gebruikt voor het uitvoeren van eenvoudige taken, is het vinden van mensen met de vereiste expertise voor het uitvoeren van complexere taken vaak een uitdaging. Nichesourcing lost dit probleem op door de expertise van bestaande groepen aan te spreken.

In hoofdstuk 4 presenteren we Accurator: een methode voor het uitvoeren van nichesourcing-campagnes, waarbij specifieke groepen betrokken worden, evenementen worden georganiseerd en een online applicatie wordt gebruikt, aangepast aan het gekozen domein. Onze bijdragen zijn: 1) een nichesourcing methodologie, 2) een annotatie-applicatie voor experts, 3) validatie van de methodologie in drie casestudies en 4) een dataset met de verzamelde annotaties. De casestudies betreffen kunstwerken met vogelafbeeldingen, bijbelprenten en afbeeldingen van mode. We vergelijken de kwaliteit en kwantiteit van de verkregen annotaties en tonen daarmee aan dat de nichesourcing methodologie in combinatie met de annotatie-applicatie gebruikt kan worden voor het verzamelen van annotaties van hoge kwaliteit in verschillende domeinen. Een gebruikersstudie toont aan dat de applicatie geschikt is voor domein-specifieke annotatietaken.

Diversificatie van zoekresultaten

In hoofdstuk 5 onderzoeken we of, en in welke mate, de verrijking van objectbeschrijvingen in de vorm van Linked Data, exploratief zoeken mogelijk maakt. Als casestudy gebruiken we de Linked Data van het Rijksmuseum, gerelateerd aan verschillende gestructureerde vocabulaires. We passen een bestaand zoekalgoritme voor graafstructuren toe op de data, welke verbanden vindt tussen een zoekterm en objecten. Daarna groepeert het algoritme soortgelijke objecten gebaseerd op de gevonden verbanden. We gebruiken het aantal gevonden verbanden en de afstand tussen zoekterm en object als indicator voor diversiteit. De zoektermen die we gebruiken zijn in de loop van een maand ingevoerd door gebruikers van de website van het museum.

De resultaten tonen aan dat binnen dit domein de verrijking leidt tot 1) een toename van het aantal zoekresultaten en 2) een toename van de variatie in zoekresultaten. Onze hypothese is dat er twee factoren invloed hebben op de geschiktheid van vocabulaires voor exploratief zoeken: 1) het aantal connecties tussen unieke concepten en objecten en 2) de rijkdom van interne connecties tussen concepten in vocabulaires. Deze vierde stap van de methode illustreert dat toegevoegde betekenis in de vorm van gestructureerde vocabulaires gebruikers kan helpen bij het verkennen van collecties en het bereiken van objecten waarin zij geïnteresseerd zijn.

Integratie van collecties

Online collecties van verschillende cultureel-erfgoedinstellingen bevatten vaak objecten die elkaar aanvullen, wat het integreren van heterogene collecties de moeite waard maakt. In

hoofdstuk 6 beschrijven we het DigiBird-systeem. Dit systeem geeft gebruikers toegang tot vier verschillende collecties en vereenvoudigt crowdsourcing initiatieven. Het systeem is ontworpen om metadata van collectieobjecten op elkaar af te stemmen, crowdsourcing bijdragen onmiddellijk beschikbaar te maken en de voortgang van crowdsourcing initiatieven in de gaten te houden op één centrale plek.

Het is uitdagend om data te harmoniseren van verschillende systemen, welke gebruik maken van een veelvoud aan standaarden en protocollen. De onderliggende data is onderhevig aan een continue stroom van crowdsourcing bijdragen. Met het DigiBird-systeem kunnen instellingen beslissen om data in een vroeg stadium van het crowdsourcing proces te gebruiken. Door het samen uitvoeren van crowdsourcing initiatieven, kunnen benodigde middelen worden gedeeld en kan er inzicht worden verkregen in de tijd die nodig is om goede resultaten te behalen. Met het gebruik van gestandaardiseerde datamodellen en annotaties uit gestructureerde vocabulaires, illustreert het DigiBird-systeem de toegevoegde waarde van verrijkingen en de voordelen van Linked Data voor collectie integratie.

Conclusie

In deze dissertatie presenteren we een methode voor het contextualiseren en verrijken van cultureel-erfgoedcollecties, om daarmee exploratief zoeken en collectie-integratie mogelijk te maken. Musea met soortgelijke collecties als het Rijksmuseum zullen de methode kunnen toepassen zonder dat grote aanpassingen nodig zijn. We verwachten dat veel van de stappen van deze methode ook in andere domeinen toepasbaar zijn. Het uitdragen van kwalitatieve informatie over objecten is onderdeel van de missie van cultureel-erfgoedinstellingen. Instellingen die rijke, gecontextualiseerde data online willen publiceren, moeten een inschatting maken van de kwaliteit van beschikbare externe vocabulaires en efficiënte methodes vinden om collectiedata te relateren aan deze nieuwe bronnen. Nichesourcing is één van deze methodes, welke instellingen in staat stelt om op een nieuwe manier het publiek te betrekken. Om mensen te motiveren is het van belang dat ze weten dat hun werk er toe doet. In onze systemen laten wij direct de invloed zien van annotaties op zoekfunctionaliteit en de mogelijkheid tot collectie integratie, waarmee we de kracht van crowdsourcing en Linked Data benadrukken.