Collection Registry
(M 1.2.2)

Version 15/06/2012
Arbeitspaket 1.2
Verantwortlicher Partner BBAW

DARIAH-DE
Aufbau von Forschungsinfrastrukturen für die e-Humanities

Dieses Forschungs- und Entwicklungsprojekt wird / wurde mit Mitteln des Bundesministeriums für Bildung und Forschung (BMBF), Förderkennzeichen 01UG1110A bis M, gefördert und vom Projektträger im Deutschen Zentrum für Luft- und Raumfahrt (PT-DLR) betreut.

SPONSORED BY THE

Federal Ministry
of Education
and Research

Projekt: DARIAH-DE: Aufbau von Forschungsinfrastrukturen für die e-Humanities
BMBF Förderkennzeichen: 01UG1110A to M
Laufzeit: März 2011 bis Februar 2014

Dokumentenstatus: Entwurf
Verfügbarkeit: DARIAH-DE-intern
Autoren:

Christoph Plutte (CP), BBAW
Patrick Harms (HP), SUB

Revisionsverlauf:

Datum	**Autor**	**Kommentar**
15/06/2012	CP	Initiale Version
26/06/2012	HP	Korrekturen
29/06/2012	CP	Korrekturen eingearbeitet

Inhaltsverzeichnis:

1. Einleitung

Die *Collection Registry* ist ein online zugängliches zentrales Verzeichnis, in dem Daten- und Forschungssammlungen registriert und Beschreibungen dieser Datensammlungen verzeichnet werden. Die *Collection Registry* bietet für die in ihr abgelegten Sammlungsbeschreibungen, die den DARIAH Metadaten-Anforderungen (siehe 3.1. Datenmodell) genügen, sowohl einen Maschinen-lesbaren als auch einen Menschen-lesbaren Zugriff. Sie ist eine zentrale Komponente in der DARIAH Infrastruktur und interagiert mit anderen DARIAH-Komponenten wie der *Schema Registry*. Eintragungen in der *Collection Registry* sind auf Sammlungsebene. Ausgehend von diesen Eintragungen können die Sammlungen auf Objektebene mittels weiterer Komponenten der DARIAH-Datenföderation zur Suche erschlossen werden.

Dieses Dokument liefert einen Überblick über den aktuellen Stand der Implementierung der *Collection Registry* und ihres gegenwärtigen Funktionsumfangs. Dabei werden sowohl die Kernaufgaben, das Datenmodell, der Funktionsumfang und die Architektur besprochen als auch mögliche Erweiterungen der *Collection Registry* diskutiert.

2. Datenförderation

2.1. Konzeptioneller Kontext

Zum Gesamtkontext der *Collection Registry* innerhalb der DARIAH-Datenföderation und in der Interaktion mit anderen DARIAH-Komponenten wie der *Schema Registry* und Suchdiensten wurden bereits in verschiedenen Dokumenten Ideen und erste Überlegungen zusammengestellt[1], zwei dieser Dokumente befinden sich im DARIAH-wiki:

- *Einordnung und Abgrenzung der Interoperabilitätsanforderungen an Collection-Registry und Schema-Registry als Grundlage für die weitere Abstimmung* (26 May 2011)

- *Modellierung semantischer Assoziationen in Forschungsdaten der Digital Humanities—Analyse der Anwendbarkeit bestehender Ansätze* (28 November 2011)

Fragen der Funktionalität und des Datenmodells der *Collection Registry* werden insbesondere in zwei Dokumenten diskutiert, die vornehmlich in der Arbeitsgruppe Daten und Sammlungen erstellt wurden:

- *Collection Registry Overview*

- *DARIAH-DE Collection Level Description Application Profile* (Mai 2012) [2]

- *Functional Requirements for the DARIAH-DE Collection Registry* (März 2012)

[1] Links zu Dokumenten, Standards und anderen Quellen finden sich im Abschnitt „Links“ am Ende des Dokuments.
[2] Wichtige Dokumente, die bisher nur im DARIAH-wiki veröffentlicht wurden, finden sich auch im Anhang dieses Dokuments.

2.2. Zusammenspiel mit anderen DARIAH-Komponenten

In der Konzeption der DARIAH-Datenföderation nimmt die *Collection Registry* eine zentrale Stelle ein, an der Datensammlungen verschiedenster Provider registriert und beschrieben werden. Für die Handhabung, Erschließung und Durchsuchbarmachung der verschiedenen Datensammlungen ist hier die Kennzeichnung der in den Sammlungen verwendeten Kodierungsstandards ebenso wie auch die Registrierung der Zugriffsdienste und –funktionen wichtig.

Da innerhalb der Datenföderation mit verschiedenen Standards gearbeitet wird und diese Heterogenität voraussichtlich auch in Zukunft bestehen wird, ist eine *Schema Registry* als zentrales Verzeichnis von Kodierungsschemata vorgesehen und wird ebenfalls im Arbeitspaket 1.2 entwickelt. Die Funktionsweise der *Schema Registry* ist im Dokument zum Meilenstein 1.2.1 beschrieben:

> *Schema Registry (M 1.2.1)*

Föderierte Suchsysteme sollen künftig mittels beider Verzeichnisse dahingehend ermöglicht werden, dass Suchanfragen auf heterogenen Datensammlungen über verschiedene Kodierungsschemata hinweg durchführbar sind. Hierfür ist eine enge Interaktion zwischen der *Collection Registry* und der *Schema Registry* erforderlich, da zu den Sammlungen auch die in ihnen verwendeten Schemata in der *Schema Registry* verzeichnet werden sollen, um mittels registrierter Crosswalks zwischen verschiedenen Schemata die Heterogenität der Daten bei der Suche zu überbrücken.

3. Collection Registry

Die *Collection Registry* umfasst drei Hauptfunktionsbereiche: Datenhaltung, Menschen-lesbare Datenanzeige und Bearbeitungswerkzeuge und eine Maschinen-lesbare Schnittstelle zu den Daten über ein standardisiertes OAI-PMH-Protokoll (*Open Archives Initiative-Protocol for Metadata Harvesting)*. Die Datenhaltung übernimmt eine relationale Datenbank, in der über verschiedene Schichten vermittelt die Datenobjekte abgelegt werden. Die Menschen-lesbare grafische Benutzerschnittstelle für die Suche und Bearbeitung von Sammlungsbeschreibungen ist als dynamische Website realisiert.

3.1. Datenmodell

Eine beliebig heterogene Sammlung von Ressourcen wird als Collection bezeichnet und beschreibt ein Konstrukt der Anwendungsdomäne, welches zur fachlichen Strukturierung von Archiven und Datenquellen eingesetzt werden kann. Collections können selbst direkt Ressourcen oder weitere untergeordnete Teilcollections beinhalten und sie aggregiert physische als auch digitale Objekte oder nur Daten. Aus diesem Grund wird die *Collection Registry* nicht einfach nur Informationen zu digital vorliegenden Ressourcen wie Texten und Bildern bereitstellen, sondern auch zu physischen Objekten und Daten. Der Zugang zu diesen Inhalten ist gleichwertig und wichtig für die Unterstützung geisteswissenschaftlicher Forschung.

Da in den Geisteswissenschaften Sammlungen auf der Grundlage völlig verschiedener Ordnungsprinzipien erstellt und gepflegt werden, ist die Erstellung eines Schemas zur Beschreibung von Sammlungen eine komplexe Aufgabe, die nur in enger Zusammenarbeit mit Fachdisziplinen vorgenommen werden kann.

Die *Collection Registry* verwendet das DARIAH Collection Level Application Profile, das auf dem Dublin Core Collection Application Profile (DCCAP) basiert, zur Beschreibung von Collections. Die Entscheidung für dieses Datenmodell auf der Grundlage von DCCAP, das bereits vielfach im Einsatz ist und gute Bedingungen für Erweiterungen bietet, wurde in der Arbeitsgruppe Daten und Sammlungen in der gemeinsamen Diskussion zwischen Fachwissenschaftlern und Informatikern getroffen. Die nähere Bestimmung des zu implementierenden Datenmodells ist in einem gesonderten Dokument der Arbeitsgruppe spezifiziert:

DARIAH-DE Collection Level Description Application Profile (Mai 2012)

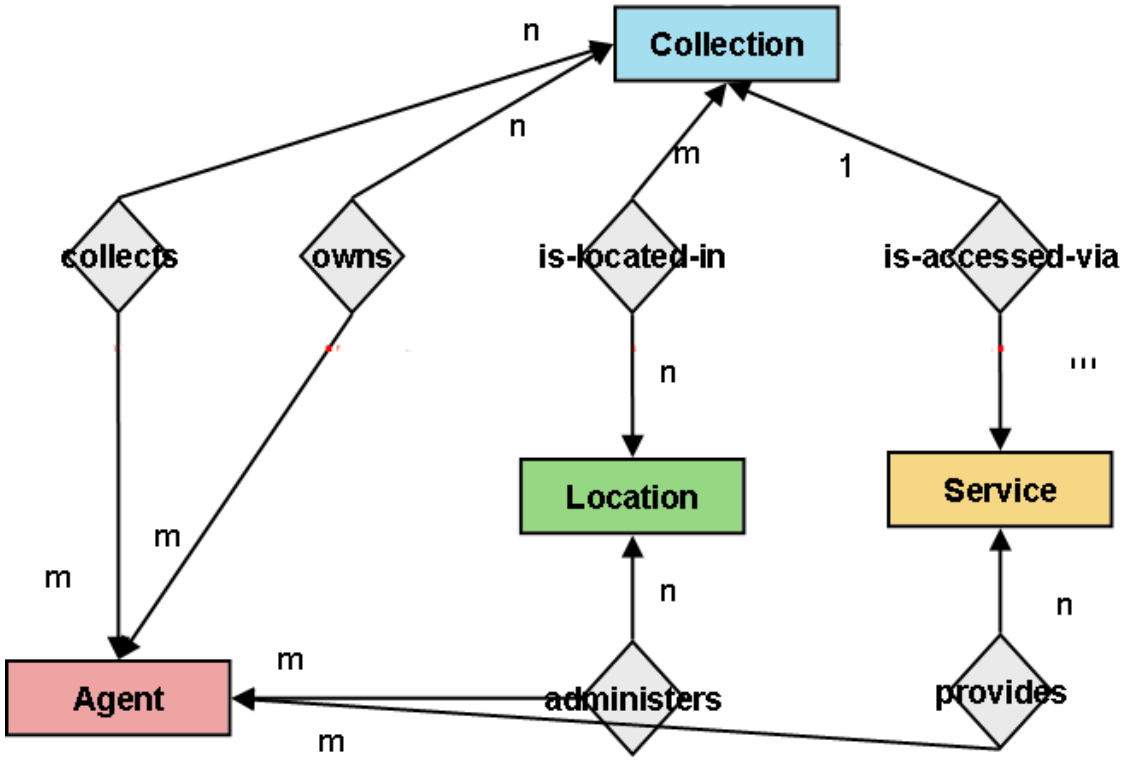


Abbildung 1: Dublin Core Collection Application Profile - Datenmodell

Die Abbildung 1 veranschaulicht das DCCAP mit seinen vier Objekten Collection, Location, Service und Agent. Die Collection-Objekte bilden den Kern des Datenmodells und repräsentieren eine Sammlung physischer oder digitaler Objekte und enthalten die Attribute zur Beschreibung der Sammlung. Die Orte, an denen eine Sammlung verwahrt wird, werden mittels des Location-Objektes beschrieben und Collections werden mittels einer is-located-in-Beziehung mit einer Location verknüpft. Die Zugriffsarten und –punkte werden als Service-Objekte beschrieben und Collections werden mit diesen durch eine is-accessed-via-Beziehung verbunden. Agent-Objekte repräsentieren sowohl natürliche Personen als auch Institutionen. Sie treten als die Ersteller bzw. Sammler von Collections auf (collects-Beziehung) und als Eigentümer von Sammlungen (owns-Beziehung). Zudem können Agent-Objekte auch als Administratoren von Location-Objekten sowie als Anbieter von Services verzeichnet werden.

Durch diese Aufgliederung der Sammlungsbeschreibung und ihrer Orte, Zugriffspunkte und Eigentümer können Redundanzen vermieden werden und Verknüpfungen einzelner Agenten mit beliebig vielen Collections etc. beschrieben werden.

Des Weiteren sieht das Datenmodell auch die Verknüpfung von Collections mit anderen Collections vor, um sie als Sub-Collections bzw. Teil einer übergeordneten Sammlung bzw. Super-Collection oder übergeordneten Sammlung zu beschreiben.

3.2. Technische Umsetzung

Die Collection Registry ist als Java Webapplication realisiert und in einer Tomcat6/JDK6/Eclipse Entwicklungsumgebung programmiert. Als Framework wurde das Spring Framework gewählt, da es mit klarer Schichtentrennung und Funktionsaspekten gute Erwei-

terbarkeit bietet. So wird die Benutzerauthentifizierung mittels Spring Security vorgenommen, das direkt für die Anbindung an den LDAP-Server von DARIAH-DE konfiguriert wurde.

Für die Datenhaltung wurde die in Spring bereitgestellte Java Persistence API-Implementierung (JPA) mit Hibernate und PostgreSQL als DBMS gewählt. Durch den weit verbreiteten object-relational Mapper (ORM) Hibernate ist die Unabhängigkeit von spezifischen relationalen DBMSs gesichert, wie auch die JPA-Schicht in Spring ebenfalls Unabhängigkeit von Hibernate als ORM bietet.

Da die Benutzerinteraktionen in der Regel in einer Reihe von Eingaben und Folgeaktionen bestehen, bot sich für die Entwicklung der GUI das Spring Web Flow-Rahmenwerk (SWF) an, in dem Interaktionsketten sinnvoll und übersichtlich konfiguriert werden können und an spätere Erweiterungen und Umstellungen des Workflows leicht angepasst werden können. Durch die Integration von Ajax ist das partielle Rendering von Formularen möglich. Zur Verbesserung der Eingabeunterstützung beispielsweise bei Datumseingaben wurden zudem Richfaces-Komponenten in die Spring-Umgebung integriert.

Die OAI-PMH bzw. RESTful-Schnittstelle ist durch eine Implementierung und Erweiterung des OAICat entwickelt. OAICat wird am Online Computer Library Center entwickelt und als konfigurierbare OAI-Content Provider Implementierung in Java mittels eines JavaServlets zur Verfügung gestellt. OAICat wurde gewählt, weil die Implementierung in Java eine einfache Einbindung in die *Collection Registry* gewährt und der Quellecode von OAICat unter Apache License Version 2.0 und damit einer von DARIAH favorisierten Lizenz verfügbar ist.

3.3. Funktionsumfang

Die *Collection Registry* bietet eine sogenannte CRUD-Schnittstelle – Create, Read, Update, Delete – als Zugang für Recherche und Bearbeitung der verzeichneten Sammlungsbeschreibungen. Recherchemöglichkeiten wie einfache Suche, erweiterte Suche und Browsen von Sammlungsbeschreibungen ist ohne Authentifizierung möglich. Für die Eingabe und Bearbeitung von Sammlungsobjekte ist eine Authentifizierung vorausgesetzt, die unten näher beschrieben wird. Da die OAI-PMH-Schnittstelle eine reine Leseschnittstelle ist, erfordert sie keine vorherige Authentifizierung.

3.3.1. GUI

Für den Lesezugriff ohne Authentifizierung bietet die GUI der *Collection Registry* drei verschiedene Suchfunktionen: einfache Suche, erweiterte Suche und Browsing. Bei der einfachen Suche wird auf allen Attributen einer Sammlungsbeschreibung nach enthaltenen Teilstrings gesucht. Bei der erweiterten Suche können beliebig viele Kriterien für verschiedene Attribute der Sammlungsbeschreibung kombiniert werden. Beim Browsing (siehe Abbildung 2) werden alle verzeichneten Sammlungsbeschreibungen berücksichtigt und in einer alphabetischen Ordnung innerhalb einer Baumstruktur zu einem gewählten Kriterium (z.B. Titel, Sachgebiet, Eigentümer) übersichtlich angezeigt. Dadurch soll die *Collection Registry* dem Benutzer ermöglichen, sich leicht einen Überblick über die vorhandene Themenabdeckung bzw. zeitliche Breite etc. der Sammlungen zu verschaffen.

Die Eingabe und Bearbeitung von Sammlungsbeschreibungen und anderen Datenobjekten ist in sogenannten Flows, d.h. Eingabeabfolgen organisiert, durch die die Benutzerin durch die Eingabe der verschiedenen Attribute und Etappen geführt wird (siehe Abbildung 3). Eine Übersicht über die bereits bearbeiteten Etappen auf der linken Seite erleichtert die Orientierung und nach jeder Etappe bzw. jedem einzelnen Eingabeformular erfolgt eine Validierung. Zu jedem Eingabefeld werden Hilfetexte beim Mouse-Over über einem Informations-Symbol angezeigt.

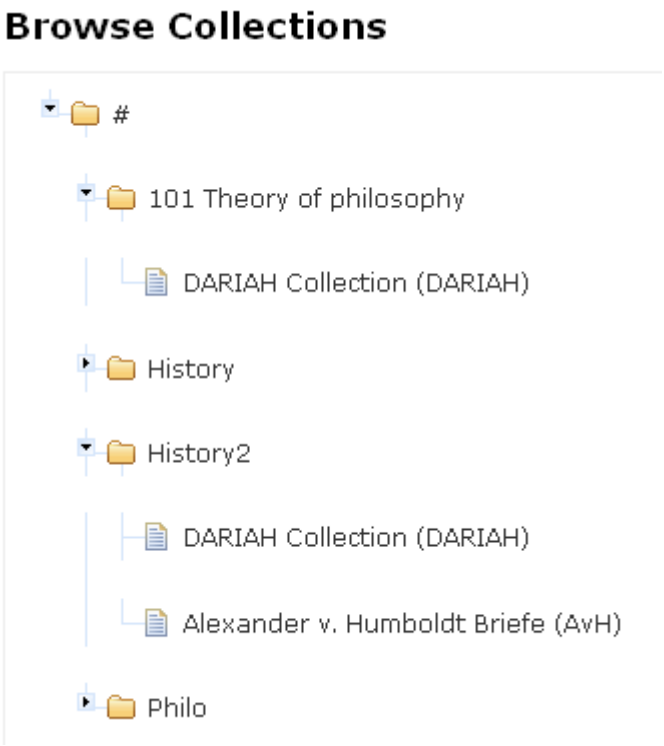


Abbildung 2: Browsing von Collections nach Themengebieten

Soweit möglich werden von der *Collection Registry* kontrollierte Vokabulare unterstützt. Für die Eingabe von Sachgebieten stehen z.B. entsprechende Eingabeunterstützungen zur Verfügung, die die Auswahl des Vokabularanbieters (z.B. Dewey, LCC) und die entsprechenden Vokabulare in einer Baumstruktur zur Auswahl anbieten (siehe Abbildung 4). Solche Eingabeunterstützungen stehen auch bei der Angabe der Sprache in den Sammlungsobjekten zur Verfügung, wobei verschiedene Kodierungen nach ISO (z.B. ISO639-3) angeboten werden.

Die *Collection Registry* unterscheidet drei verschiedene Benutzer: nicht authentifizierte Benutzer ohne Schreibrechte, die Daten durchsuchen und anzeigen können, authentifizierte Benutzer aus der Benutzergruppe „collection-registry-users“ und authentifizierte Benutzer aus der Gruppe „collection-registry-admins“.

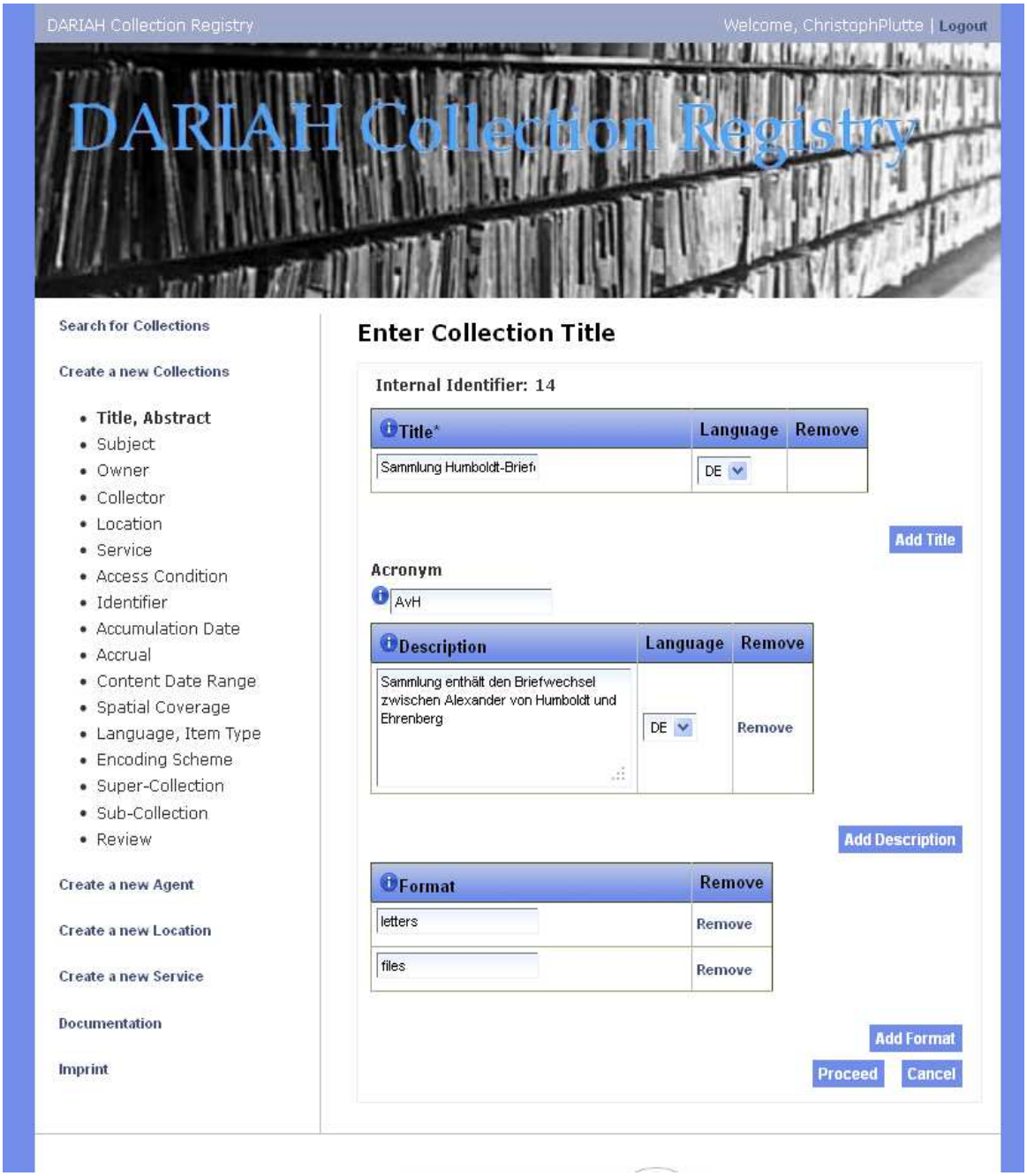


Abbildung 3: Formular zur Eingabe von Titel und Beschreibung.

Mitglieder der Gruppe „collection-registry-users“ dürfen neue Datenobjekte anlegen und die von ihnen angelegten Objekte bearbeiten und löschen. Jedes Datenobjekt besitzt ein Reviewed-Attribut, das zunächst auf „false“ gesetzt wird. Benutzer der ersten Gruppe dürfen dieses Attribut nicht bearbeiten, sie können so zwar Sammlungsbeschreibungen vorschlagen, diese sollen dann aber noch einen Review-Prozess durchlaufen.

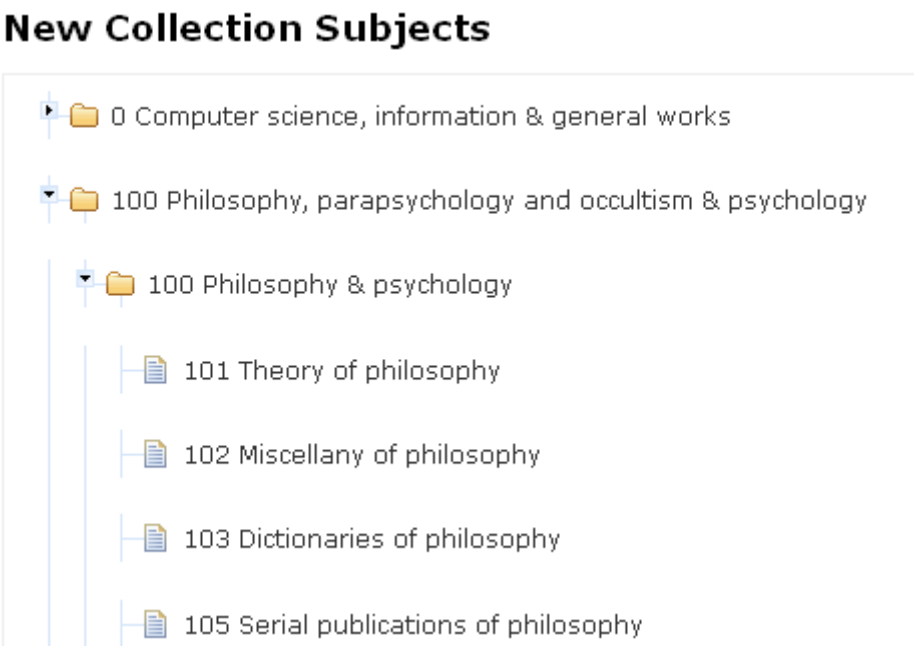


Abbildung 4: Dewey Dezimalklassen als Auswahlbaum

Die Gruppe „collection-registry-admins“ umfasst die Administratoren der *Collection Registry*, die über dieselben Rechte wie die erste Gruppe verfügen mit dem Unterschied, dass sie nicht nur die von ihnen selbst angelegten Objekte bearbeiten und löschen dürfen, sondern ebenfalls die Objekte anderer Benutzer. Außerdem dürfen Administratoren das Reviewed-Attribut bearbeiten und somit die Korrektheit einer Sammlungsbeschreibung festhalten.

Die Authentifizierung und Autorisierung der Benutzer in der *Collection Registry* erfolgt über die DARIAH-AAI. Da es aufgrund verschiedener offener Fragen im Zusammenhang mit der AAI zum Zeitpunkt der Abfassung dieses Dokuments noch nicht möglich war, die Shibboleth-Infrastruktur in die *Collection Registry* zu integrieren, wurde zunächst eine Anbindung an den DARIAH-DE-LDAP-Server vorgenommen. Die Authentifizierung sowie die Autorisierung erfolgen gegenwärtig über diesen Server. Auch nach der Einbindung der Shibboleth-Infrastruktur für die Benutzerauthentifizierung ist es geplant, die Autorisierung, d.h. die Benutzergruppen- und –rechteverwaltung über den DARIAH-DE-LDAP-Server vorzunehmen.

3.3.2. OAI-Schnittstelle

Als Maschinen-lesbare Schnittstelle bietet die *Collection Registry* eine OAI-PMH-Schnittstelle für Metadaten-Harvesting. Die Schnittstelle ist wie oben beschrieben als Java-Servlet realisiert und über eine URL erreichbar. Es sind die üblichen OAI-PMH-Verbs und Parameter für folgende Funktionen implementiert:

- Identify: Gibt die Identifizierung des OAI-Providers der *Collection Registry* zurück.
- GetRecord: Liefert einen einzelnen Eintrag der *Collection Registry*.
- ListRecords: Liefert eine Liste der vorhandenen Einträge.

- ListIdentifiers: Liefert eine Liste der vorhandenen Identifikatoren von Sammlungen.

- ListMetadataFormats: Liefert eine Liste der angebotenen Metadatenformate.

Für die Serialisierung der Sammlungsbeschreibungen in XML für die Ausgabe über die OAI-Schnittstelle stehen zwei Formate zur Verfügung: oai_dc und dclap. oai_dc ist die standardisierte XML-Serialisierung von Dublin Core für die OAI-PMH-Schnittstelle und ist der am weitesten verbreitete Standard für OAI-PMH. Da oai_dc nur Dublin Core Elemente und keine Dublin Core Erweiterungen wie dcterms oder DCCAP enthält, wurde dclap als eigener Standard entworfen.

DCLAP bzw. DARIAH Collection Level Application Profile ist das spezifisch für die *Collection Registry* entwickelte Austauschformat für die OAI-Schnittstelle. Dieses Format umfasst alle Attribute der Sammlungsbeschreibungen einschließlich der mit ihnen verknüpften Objekte wie Location, Service und Agent.

Das DCLAP ist mit einem XML-Schema beschrieben und auf einem DARIAH-Server hinterlegt.

4. Ausblick

An dieser Stelle werden als Ausblick für künftige Entwicklungen Erweiterungen diskutiert, die in den Ausbau und die Weiterentwicklung der *Collection Registry* einfließen können.

Internationalisierung: Geplant ist die Internationalisierung der GUI für Deutsch, Englisch und ggfs. weitere Sprachen aus dem Kreis der DARIAH-EU-Mitglieder. Bestandteil der Internationalisierung sollten nicht nur die Bezeichnungen von Links und Buttons sein, sondern auch der Hilfetexte, Dokumentationen sowie nach Möglichkeit der kontrollierten Vokabulare. Da die letzten Punkte jedoch einen gewissen Zeitaufwand erfordern und die Entscheidungen über die zu verwendenden Vokabulare zum Zeitpunkt der Abfassung noch nicht getroffen waren, konnte die Internationalisierung bisher nur vorbereitet, aber noch nicht abgeschlossen werden.

Layout und Design: Der Webauftritt der *Collection Registry* ist aktuell in einem vorläufigen Design gehalten. Zu überlegen wäre hier, in welches Layout der Webauftritt langfristig gebracht werden soll und wo Entwürfe für eine Überarbeitung des Designs entstehen können, damit der Webauftritt der *Collection Registry* dahingehend angepasst werden kann.

Shibbolethisierung: Wie oben bereits besprochen ist eine Anbindung an die Shibboleth-Infrastruktur und damit die vollständig Einbindung in die DARIAH-AAI geplant. Künftig soll die Benutzerauthentifizierung über Shibboleth abgewickelt werden, die Autorisierung, d.h. die Zuordnung eines Benutzers zu bestimmten Benutzergruppen bzw. –rollen wird dann ggfs. über den DARIAH-LDAP-Server geleistet. Wie beide Systeme für Authentifizierung und Autorisierung zusammenarbeiten war zum Zeitpunkt der Abfassung noch nicht ganz entschieden, sobald hier eine langfristige Strategie vorliegt, kann die Shibbolethisierung der *Collection Registry* erfolgen.

Zusammenführung von Collection Registry und *Schema Registry*: Um das oben beschriebene Zusammenspiel von *Collection Registry* und *Schema Registry* zu erleichtern und zu verbessern ist eine Zusammenführung und Vereinigung beider Systeme in einem einheitlichen Webportal geplant. Voraussetzung ist hierfür die Fertigstellung der *Schema Registry*, die sich zum Zeitpunkt der Abfassung noch in einem prototypischen Stadium befand.

Review-workflow: Zur komfortableren Unterstützung des moderierten Ansatzes und des Review-Prozesses und der Qualitätskontrolle der vorgeschlagenen bzw. verzeichneten Sammlungsbeschreibungen kann eine entsprechende Workflow-Unterstützung entwickelt werden. Da zum Zeitpunkt der Abfassung die Anforderungen an einen solchen Prozess noch nicht ausreichend spezifiziert waren, kann diese Unterstützung erst entwickelt werden, wenn hier erste Erfahrungen und Anforderungsbeschreibungen vorliegen.

Links

Links wurden zuletzt am 15. Juni 2012 aufgerufen.

Dokumente

Collection Registry Overview ;
https://dev2.dariah.eu/wiki/display/DARIAHDE/Collection+Registry+Overview

DARIAH(-DE): Digital Research Infrastructure for the Arts and Humanities—Concepts and Perspectives;
http://collab.teldap.tw/teldap2012/

Einordnung und Abgrenzung der Interoperabilitätsanforderungen an Collection-Registry und Schema-Registry als Grundlage für die weitere Abstimmung (conceptual draft);
https://dev2.dariah.eu/wiki/download/attachments/2295602/AP1_2_Interoperabilit%C3%A4t.pdf?version=1&modificationDate=1326219424816

DARIAH-DE Collection Level Description Application Profile;
https://dev2.dariah.eu/wiki/display/DARIAHDE/DARIAH-DE+Collection+Level+Description+Application+Profile

Functional Requirements for the DARIAH-DE Collection Registry;
https://dev2.dariah.eu/wiki/display/DARIAHDE/Functional+Requirements+for+the+DARIAH-DE+Collection+Registry

Schema Registry (M 1.2.1);
https://dev2.dariah.eu/wiki/download/attachments/2295237/M1.2.1+Schema+Registry-2.pdf?version=1&modificationDate=1334923337614

Shibboleth und DARIAH-AAI;
https://dev2.dariah.eu/wiki/download/attachments/2295732/DARIAH-AAI-Concept-v0.3a.pdf?version=1&modificationDate=1337876106997

Genannte und verwendete Standards

Apache License Version 2.0;
http://www.apache.org/licenses/LICENSE-2.0

Dublin Core (DC);
http://dublincore.org

Dublin Core Collection Application Profile;
http://dublincore.org/groups/collections/collection-application-profile/

DCTERMS;
http://dublincore.org/documents/dcmi-terms/

Open Archives Initiative-Protocol for Metadata Harvesting;
http://www.openarchives.org/pmh/

OAI_DC;
http://standards.jisc.ac.uk/catalogue/OAI_DC.phtml

Frameworks und Komponenten

Apache Tiles;
http://tiles.apache.org/

Apache Tomcat;
http://tomcat.apache.org/

Eclipse;
http://www.eclipse.org/

Hibernate;
http://www.hibernate.org/

Java Development Kit (JDK);
http://www.oracle.com/technetwork/java/javase/overview/index.html

OAICat
http://www.oclc.org/research/activities/oaicat/default.htm

PostgreSQL;
http://www.postgresql.org/

Richfaces
http://www.jboss.org/richfaces

Spring 3;
http://www.springsource.org/

Spring Web Flow
http://www.springsource.org/spring-web-flow

5. Anhang

DARIAH-DE Collection Level Description Application Profile

Hinzugefügt von Wibke Kolbmann , zuletzt bearbeitet von Christoph Plutte am Jun 06, 2012

Introduction – Collection Description in the Arts and Humanities

This document describes the DARIAH-DE Collection Level Description Application Profile, a schema designed for the DARIAH-DE Collection Registry. The Collection Registry is a web application to allow users register collections for the DARIAH-DE infrastructure giving access to digital resources which are of research interest in the arts and humanities. Furthermore the collection registry will help the community to take informed decisions in collection management. The resources will be made available for discovery and reuse for research purposes only.

A collection aggregates physical or digital items or simply data facts. The DARIAH-DE Collection Registry therefore will collect and make accessible information not only on digital available resources like texts or images but also from physical collections and data facts since the access to all these resources is considered to be of importance to support research in the arts and humanities.

The definition of a schema for collection description is challenging due to the ordering of collections based on varying principles in the different arts and humanities. George Macgregor defined a collection level description ‘to be a structured, open, standardised and machine-readable form of metadata providing a high-level description of an aggregation of individual items. Such descriptions disclose information about their existence, characteristics and availability, and employ the use of implicit item-level metadata and, more particularly, contextualise that aggregation of item-level descriptions.#’

In general there is a need for collection level descriptions to apply a minimum standard of documentation to collections of resources, if there are not enough resources for cataloguing on item level.

Collection descriptions contain information which help institutions to manage their collection, for example regarding backup and maintenance of data and resources or for curatorial responsibilities like collection development, status of digitization of collection items, status of indexing, storage management, preservation or need for migration, aggregation of collections. They should enable institutions to take informed strategic decisions at institutional, cross-institutional, regional, sectoral and national levels.

To enable reuse of research data and resources collection descriptions allow for discovery, selection and querying across holding of diverse institutions and databases, furthermore making resources maintained by one party accessible for another party to develop a new application or service on.

The mission of DARIAH is to enhance and support digitally-enabled research across the humanities and arts. DARIAH aims to develop and maintain an infrastructure in support of ICT-based research practices. The need for collection level description information particularly for DARIAH-DE therefore focuses on:

- enabling interfacing data repositories with services and applications,
- improving accessibility of sensitive data for research purposes within a trusted registry due to differentiated access rights,
- monitoring the growth of the partner network,
- monitoring of usage of offerings of the infrastructure to estimate further needs for digitization and technical development

The definition of the DARIAH-DE Collection Level Description Application Profile is based on the following principles:

- to allow a broad diversity of information since it is easier to ignore given information than to collect and integrate information afterwards
- to allow customisation of the data model
- to keep the barriers for registration for the Collection Regsitry low by leaving most properties optional
- to define a small group of core properties mandatory ensuring a minimum standard of quality data for reuse
- to focus on interoperability with other aggregation platforms of research data in the arts and humanities on a national and European level, therefore also considering multilingual issues
- to allow for reuse of collection description metadata in other standards

The DARIAH-DE Collection Description Application Profile builds on these standards and the work of these members of the DARIAH-DE partner institutes: Dublin Core Collection Application Profile (DCCAP), vCard,Thomas Kollatz, Stephan Schmunk, Kristin Herold, Matteo Romanello and Wibke Kolbmann, Niels-Oliver Walkowski.

The discipline - specific Perspective – Definition of collection, research

interest, collection description in discipline - specific standards

In general collection information to a researcher is more interesting regarding creator, subject, type, time or spatial coverage of a collection of single items. Information on location, uniqueness or ownership of a collection, which is more important to archives, libraries and museums for collection management, is finally important to access an item of research interest but it is less of importance to the discipline-specific research itself. Therefore special needs for collection descriptions exist coming from different disciplines in the arts and humanities that are outlined in the next sections.

Archaeology

The history of archaeology is itself of interest for research of archaeologist. It investigates the changes in techniques and increase in professionalism of the science. Collections of antiquities were the first step towards the foundation of archaeology as a science. Of interest for archaeologists here is when, where and by whom these collections were gathered and how they specialized on specific items like coins, vases or artefacts of a specific culture (Etruscan, Egyptian etc.), further - what was the intention and purpose of the collector to select and gather archaeological objects. Later universities started to build plaster cast collections# of antiquities for study purposes, producing replica from originals. Many of these in the beginning private collections are now part of public museums#.

Collections in an archaeological context can also be understood as the sum of all artefacts, monuments or other findings of an excavation place. To protect archaeological sites and artefacts the access to the information about the places and items is restricted.
Sub-collections could be built on all fragments of a building for example.

With the beginning of the digitization surrogates of archaeological objects or the documentation of excavations were entered into separate databases for special media types: graphics, text, 2D/3D, maps (GIS). The challenge today is to link these resources together again and recontextualize the findings in a virtual space.

There are some archaeology specific metadata standards and others from the cultural heritage domain that contain collection level description information particularly interesting for archaeologists. As there are: CARARE, MICHAEL-EU, Adex, CDWA, VRA Core, ArchaeoML, LIDO, EDM.

Codicology

In codicology the concept of “Collection Level Description” is well known and used apart from its meaning in metadata jargon. Documents such as manuscripts are usually grouped into collections and scholars studying manuscripts are interested not only in the history of single resources but also in that of a collection as a whole (who created the collection, when was it donated to a given library or archive, etc.). Information that can be found in a collection level description for manuscripts include:

- Reference
- Title
- Date of creation
- Extent
- Language of Material
- Administrative/Biographical History
- Scope and Content
- Administrative Information
    - Immediate source of acquisition
    - Access conditions
- Finding aids
- Access points

As an example see a CLD from the Bodleian Library in Oxford.

Epigraphy

There is no CLD designed for epigraphic data in particular. Nevertheless there are collections in epigraphy. A collection could be defined as a dataset (single inscription) connected to a specific location (street, cemetery, town, etc.). Sometimes it is possible to assign the collection to a limited time span.

EpiDoc - Epigraphic Documents in TEI XML provides an interdisciplinary format for epigraphic objects see (http://epidoc.sourceforge.net/) and has a RNGSchema: http://www.stoa.org/epidoc/schema/latest/tei-epidoc.rng for data, but TEI does not support CLD.

History

Encoded Archival Description (EAD) +
Encoded Archival Context (EAC)

“The EAD metadata schema provides an XML encoding for archival descriptions. It adopts a multi-level approach to description, providing information about a collection as a whole and then breaking it down into groups, series and (if significant) individual items. EAD grew out of work done at UC Berkeley in the mid 1990s and was influenced by TEI and ISAD(G) (see below). Version 1.0 was released in 1998 with a major revision in 2002 (Version 2002). EAD is maintained by the US Library of Congress and Society of American Archivists, but is used internationally, including the UK. The DACS content standard (see above) provides guidelines for US archivists on how to enter data into EAD.”#

Documents:

- http://www.bundesarchiv.de/imperia/md/content/daofind/ead_anwenderleitfaden.pdf (Anwenderleitfaden Bundesarchiv)
- http://www.bundesarchiv.de/imperia/md/content/instada/fox.pdf (PPP)
- http://webdoc.gwdg.de/edoc/p/fundus/4/pitti.pdf

- http://www.loc.gov/ead/
- http://www.archivists.org/saagroups/ead/
- http://xml.coverpages.org/ead.html
- http://www3.iath.virginia.edu/eac/

Musicology

Collections in Musicology consist of different materials and types, for example collections of:

- musical instruments
- notes
- bibliographies
- sound carrier
- digitized notes (e.g. MEI, MusicXML)

Single items of a collection are often described with a lot of details and they are summarized in a schema, for example see RISM# (Repertoire International des Sources Musicales).

That kind of practice is not very popular on the collection level in the musicology. The collections are described with elements and facts, but the information is given in a verbal formulation, instead of a collection level description. Nevertheless some institutions, mostly libraries, are using CLDs for their collections. Here are three examples in a musicology context:

Music Manuscripts from the Library of St. Michaels´s College, The Bodleian Library of the University of Oxford
Papers of Granville Bantock, The Archives Hub
Instrumentalmusik der Dresdner Hofkapelle, Michael-DE

There is a virtual library of musicology, the VIFA MUSIC (Virtuelle Fachbibliothek Musikwissenschaft). In the digital library section (Digitization of musical works - displayed in General Collections, Thematic Collections, Countries & Regions, People), digital collections are listed and linked, but without any collection description.

Literature/Philology

Literary scholars are probably more interested in resource-level rather than collection-level descriptions. Or at least the metadata they use to describe single resources are more detailed than those they use to describe collections.
As an example, we can consider the collections contained in the Perseus Digital Library#. Although the metadata are not made explicit, what matters about a collection within the digital library is:

- collection title
- language of the texts contained in the collection
- time span covered by the texts
- creator of the collection (may be a user, another institution, etc.)
- number of items in the collection

For collections of texts these fields seem to be sufficient.

Collection Level Descriptions – Application Profiles of European Partners in DARIAH-EU aggregating research data in the arts and humanities

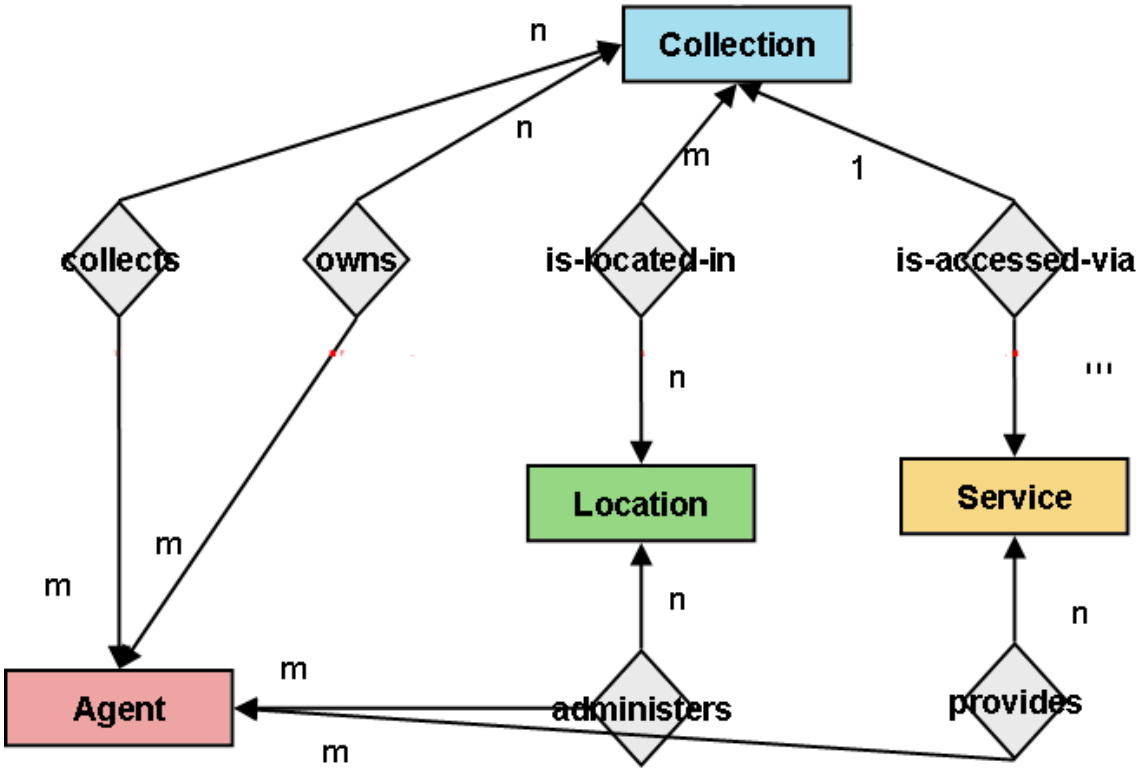
EASY.nl

ISIDORE
DHO:Discovery

DARIAH-DE Collection Level Description Application Profile Specifications

Das DARIAH-DE Collection Level Description Application Profile folgt weitestgehend dem Klassenmodell von Dublin Core Collection Application Profile (s. u.). Das Datenmodell sieht 4 Klassen vor von denen eine (Collection) die Beschreibung der Sammlung selbst darstellt. Die Klasse Agent erfasst mit der Sammlung in Verbindung stehende Personen oder Organisationen. Die Klasse Location beschreibt physische oder virtuelle Orte in denen die beschriebene Sammlung beherbergt ist. Die Service Klasse ist dazu da eine Beschreibung von Schnittstellen zu ermöglichen mit denen auf die Sammlung zugegriffen werden kann. Ein Diagramm zur übersichtlichen Darstellung der Klassen und ihrer Verbindungen zueinander wird im folgenden gegeben:

DARIAH-DE CLDAP Class Diagram



Prädikate und Vokabulare

Entsprechend der Klassen des DCLDAP wird in den folgenden Tabellen das Inhaltsmodell und die Prädikate der jeweiligen Klasse aufgelistet und beschrieben. Jedes Feld kann über die gemachten Angaben hinaus in unterschiedlichen Sprachen verwendet werden, da jedem Feld ein Sprachattribut zugeordnet ist. Ebenfalls gibt es für jedes Feld ein Scheme-Attribut indem festgehalten wird welchem Vokabular der Wert eines Feldes entnommen ist. Bei, von der Collection Registry vorgegebenen Vokabularen wird der Name des Vokabulars automatisch eingetragen. Wo dies nicht der Fall ist oder vom Benutzer zusätzliche Vokabulare über das Userinterface hinzugefügt werden können (siehe “u. a.”) wird die Angabe vom Benutzer übernommen. Eine solche Regelung lässt es z. B. zu bei dcterms:spatial neben Ortsnamen auch Koordinaten eines bestimmten Koordinatensystems festzuhalten oder eine disziplinspezifische Verschlagwortung hinzuzufügen.

Vokabulare

Vocabulary Title	Namespace Name	Prefix
The Dublin Core Metadata Element Set, v1.1	http://purl.org/dc/elements/1.1/	dc
Dublin Core Terms	http://purl.org/dc/terms/	dcterms

Dublin Core Type Vocabulary	http://purl.org/dc/dcmitype/	dcmitype
MARC Relator Code Properties	http://www.loc.gov/loc.terms/relators/	marcrel
Collection Description Terms	http://purl.org/cld/terms/	cld
Collection Description Type Vocabulary Terms	http://purl.org/cld/cdtype/	cdtype
DCLAP	-	dclap
VCard	http://www.w3.org/2006/vcard/ns#	vcard
ISO639-3	http://www.sil.org/iso639-3/	-
ISO8601	http://purl.org/dc/terms/ISO8601	-
IANA Mime Types	http://purl.org/NET/mediatypes (http://mediatypes.appspot.com/)	imt
Dewey Dezimal Klassifikation	http://purl.org/NET/decimalised#d	ddc
Library of Congress Subject Headings	http://id.loc.gov/authorities/subjects	lcsh
Geonames	http://www.geonames.org	gn
Getty Thesaurus of Geographical Names		dc:TGN

Vokabular für die Auswahllist bzgl. des Lizenztyps:
Berkeley Database License, Creative Commons (in allen Varianten), ODbL, ODC-By, Andere

Collection Klasse

Erläuterung zu den Abkürzungen: M=Pflichtfeld, O=Optional, u. a. = und andere Vokabulare, die durch den Benutzer eingetragen werden können, 1=maximal einmal zu verwenden, 1+=ein- oder mehrmals zu verwenden.

Feld	Prädikat	Beschreibung	Werte	Status
Titel	dc:title	Titel der Sammlung	Freitext	M1
Akronym	dclap:acronym	Akronym	Freitext	O1
ID	dc:identifier	Eindeutige Identifikatoren	ColReg (intern); u. a.	M1+
Typ	dc:type	Beschreibungstyp	“Collection”	M1
Beschreibung	dcterms:abstract	Eine Beschreibung der Sammlung	Freitext	O1
Größe	dcterms:extent	Die Größe der Sammlung (Objekte, MB …)	Freitext	O1
Sprache	dc:language	Sprache der Objekte in der Sammlung	ISO639-3	O1+
Objekttyp	cld:itemType	Typ der Objekte in der Sammlung nach DCMIType	DCMIType	O1+
Objektformat	cld:itemFormat	physisches und/oder digitales Format der Objekte in der Sammlung	IANA MimeTypes; u. a.	O1+
Rechte	dc:rights	Rechte	Lizenzen	O1
Zugangsbeschränkungen	dcterms:accessRights	Rechte bzgl. des Zugangs zur Sammlung	Freitext	O1
Erwerbsmethode	dcterms:accrualMethod	Methoden mittels derer Objekte zur Sammlung hinzugefügt werden	Freitext	O1
Aktualisierungshäufigkeit	dcterms:accrualPeriodicity	Frequenz innerhalb derer die Sammlung aktualisiert wird	Freitext	O1

Richtlinien	dcterms:accrualPolicy	Richtlinien für der Erstellung der Sammlung	Freitext	O1
Sammlungsgeschichte	dcterms:provenance	Besitztumswechsel und andere relevanten Herleitungsinformationen	Freitext	O1
Zielgruppe	dcterms:audience	Personenkreise für die die Sammlung von Interesse	Freitext	O1+
Schlagwort	dc:subject	Schlagwörter für die Sammlung	DDC; LCSH; +	O1+
geographische Relevanz	dcterms:spatial	Bezeichnungen oder Koordinaten für die geographische Relevanz der Sammlung	GeoNames; Getty; u. a.	O1+
zeitliche Relevanz	dcterms:temporal	Bezeichnungen oder Zeitangaben für die zeitliche Relevanz der Sammlung	ISO8601	O1+
Sammelzeitraum	dcterms:created	Zeitraum innerhalb dessen die Objekte der Sammlung zusammengetragen wurden	ISO8601	O1+
Datierungszeitraum	cld:dateItemsCreated	Zeitraum innerhalb dessen die Objekte der Sammlung erstellt wurden	ISO8601	O1+
Sammler	dc:creator	Person oder Organisation, die als Sammler der Zusammenstellung fungiert	AGENT	O1+
Besitzer	marcrel:own	Person oder Einrichtung, die als Rechteinhaber der Sammlung fungiert	AGENT	M1+
Ort der Sammlung	cld:isLocatedAt	Physischer oder virtueller Ort der die Sammlung hostet	LOCATION	M1+
Zugang zur Sammlung	cld:isAccessedVia	Schnittstelle über die auf eine Sammlung zugegriffen werden kann	SERVICE	O1+
Sub-Sammlung	dcterms:hasPart	Teilsammlung der vorliegenden Sammlung	COLLECTION	O1+
Super-Sammlung	dcterms:isPartOf	Sammlung deren Teil die vorliegende Sammlung ist	COLLECTION	O1
Metadatenschemata	dclap:metadataEncodingScheme	Das Schema, welches für die Metadaten der Objekte der Sammlung verwendet wurde	Freitext	O1+
Objektschemata	dclap:itemEncodingScheme	Schema welches in den Objekten der Sammlung verwendet wird	Freitext	O1+

Agent Klasse

Feld	**Prädikat**	**Beschreibung**	**Werte**	**Status**
Titel	dc:title	Name des Agente	Freitext	M1
Typ	dc:type	Klassentyp	“Person|Organisation”	M1
ID	dc:identifier	Persistente Identifikator	ColReg (intern)	M1+
Acronym	vcard:nickname	“		O1
Straße	vcard:street-address	“		O1
Post Box	vcard:post-office-box	“		O1
Lokalität	vcard:locality	“		O1
Postleitzahl	vcard:postal-code	“		O1
Region	vcard:region	“		O1
Land	vcard:country-name	“		O1
Telefon	vcard:Tel	“		O1

Handy	vcard:Cell	“		O1
Fax	vcard:Fax	“		O1
Email	vcard:Email	“		O1
Homepage	vcard:url	“		O1

Location Klasse

Feld	Prädikat	Beschreibung	Werte	Status
Titel	dc:title	Titel der Sammlung	Freitext	M1
Typ	dc:type	Klassentyp	“Location”	M1
ID	dc:identifier	Persistenter Identifikator		M1+
Acronym	vcard:nickname	“		O1
Straße	vcard:street-address	“		O1
Post Box	vcard:post-office-box	“		O1
Lokalität	vcard:locality	“		O1
Postleitzahl	vcard:postal-code	“		O1
Region	vcard:region	“		O1
Land	vcard:country-name	“		O1
Telefon	vcard:Tel	“		O1
Handy	vcard:Cel	“		O1
Fax	vcard:Fax	“		O1
Email	vcard:Email	“		O1
Homepage	vcard:url	“		O1
Administrator	marcrel:consultant	Ansprechpartner für den beschriebenen Ort	AGENT	O1+

Service Klasse

Feld	Prädikat	Beschreibung	Werte	Status
Titel	dc:title	Name der Schnittstelle	Freitext	M1
ID	dc:identifier	Persistenter Identifikator	ColReg (intern); u. a.	M1+
Typ	dc:type	Klassentyp	“Service”	M1
Acronym	dclap:acronym	Abkürzung für die Schnittstelle	Freitext	O1
Service Access Method	dclap:serviceAccessMethod		Freitext	M1
Service Function	dclap:serviceFunction		Freitext	M1
Service URL	dclap:serviceURL	Basis URL der Schnittstelle	URL	M1
Service Interface	dclap:serviceInterface		Freitext	O1
Service Help URL	dclap:serviceHelpURL	URL der Dokumentation für die Schnittstelle	URL	O1
Access Control	dclap:accessControl	maschinenlesbare Information bzgl. des Zugangs zur Sammlung	MARC 506	M1
Administrator	marcrel:consultant	Person die die Schnittstelle administriert	AGENT	O1+

Beispiel

Im folgenden soll das DCLDAP am Beispiel von Arachne des Deutschen Archäologischen Instituts demonstriert werden:

Feld	Wert
Type	Collection
Collection Identifier (1 internal + opt.)	6546547584765
Title	Arachne
Description	Arachne is the central Object database of the German Archaeological Institute (DAI) and the Archaeological Institute of the University of Cologne (for further info see Arachne's homepage). It contains DAI's collection of images of archaeological objects and books (iDAI.BookBrowser) related to Classical Archaeology.
Size	836.142 digital Objects of 250.000 physical Objects
Language	lat, grc
Item Type	Image, Physical Object, Text
Item Format	Mosaik, Malerei, Vasen, Plastik, Idealplastik, Porträtplastik, Sarkophage, Reliefs, Figürliche Bauplastik, Inschriften, Stuck, Terrakotta, Bronze, Kleinkunst, Denkmäler, Architektur
Item Format @scheme=IANA	image/jpeg, pdf
Rights	Some contents are licensed under Creative Commons, others are not.
Access Rights	The Access to the content via webinterface is not restricted
Accrual Method	-
Accrual Periodicity	-
Accrual Policy	-
Custodial History	-
Audience	Classicists, Archaeologists
Subject @scheme=LCSH	Archeology, Photography in archaeology, …
Subject @scheme=DDC	930
Spatial Coverage @scheme=Getty	Hellas, …
Temporal Coverage	-21000/+1400
Dates Collection Accumulated	-
Dates Items Created	-
Relationships between the Collection and Agents	
Collector	AGENT ID, http://d-nb.info/gnd/160680751
Owner	AGENT ID, http://d-nb.info/gnd/160680751
Relationships between the Collection and Location, Collection and Service	
Is Located At	LOCATION ID
Is Accessed Via	SERVICE ID
Relationships between Collections (and between Collections and Catalogues or Indices)	
Sub-Collection	-

Super-Collection	-
Additional Fields of the DARIAH Collection Registry	
Metadata Scheme	CIDOC-CRM, Dublin Core, METS/MODS, Arachne XML
Item Scheme	jpeg, pdf, TEI
Agent	
Title	Reinhard Foertsch
Type	Person
Identifiers @scheme=PND	http://d-nb.info/gnd/160680751
Acronym	
Street	
Post BOX	
Locality	
Post Code	
Region	
Country	
Phone	
Mobile Phone	
Fax	
Email	
Homepage	http://www.arachne.uni-koeln.de/
Service	
Title	Arachne OAI-PMH Schnittstelle
Identifier	
Type	Service
Acronym	OAI-PMH
Service Access Method	OAI-PMH
Service Function	READ
Service URL	http://arachne.uni-koeln.de:8080/OAI-PMH/oai-pmh.xml
Service Interface	
Service Help URL	http://www.arachne.uni-koeln.de/drupal/?q=de/node/234
Acces Control	Open
Administrator	AGENT