Institutions such as libraries, museums and other archives have been collecting books, paintings, statues and other objects for centuries. To manage these collections, cataloguers have described each object with respect to its title, author, subjects, materials and other attributes. This process is called “indexing”, and simplifies the process of searching through the collections. An object description created during indexing is essentially a set of attribute-value pairs. Such a description might consist e.g. of pairs author=Rembrandt, title=Anatomy Lesson, date=1632, type=painting, subject=group portrait. Such descriptions are also called metadata (data about the actual object).

The values for the pairs are usually taken from vocabularies. Vocabularies are lists of concepts with definitions and play a key role in indexing and search. Firstly, they offer a set of agreed upon concepts that cataloguers can pick from. Secondly, concepts provide a convenient place to group synonymous terms (e.g. “clair-obscure” and “chiaroscuro” which both refer to Rembrandt’s painting style). Thirdly, the concepts usually have a unique identifier, which allows the cataloguer to indicate the correct concept even though the concept has an ambiguous term (e.g. “painting” as the process of applying a protective coating to an object vs. “painting” as the process of creating an expressive or communicative image). Fourthly, the concepts are often placed into a hierarchy (e.g. “origami” below “Japanese art”) which simplifies search (a search for books on Japanese art will also return books on origami).

To aid the indexing process, each institution not only prescribes a set of vocabularies to be used, but also the attributes and their names. Attributes are called elements, and the set of allowed attributes is together called the metadata element set.

With the advent of the Web people and institutions have started sharing their data. This has the potential benefit that all information on a particular topic, say the paintings of Vincent van Gogh, can be queried as if they were stored in one system. This requires that the data is integrated: it must conform to a particular format and structure that the search system understands. Two obstacles to integration are the different data formats in use (the syntactic integration problem) and the different terms in use to denote similar concepts (the semantic integration problem). One example of the latter problem is when one institution has a concept called “clair-obscure” and another has a concept called “chiaroscuro”. Another example is when one institution uses a metadata element called “creator” while another has an element called “author”. These concepts and elements are highly similar and a search for clair-obscure paintings by Rembrandt needs to query through both concepts and elements.

The Semantic Web is a research area that proposes particular solutions to these problems. Firstly, it proposes to use a family of web-based knowledge representation languages that have
RDF as underlying model (RDFS and the several flavours of OWL). Conversion of data to this family of languages solves a substantial part of the syntactic integration problem. Secondly, the languages have a few simple mechanisms to relate similar concepts to each other, solving part of the semantic integration problem. For example, it is possible to state that “creator” is equivalent to “author”, so that a query on either element will automatically include the results obtained from querying with the other.

In this thesis we assume that the approach and languages proposed by the Semantic Web community are useful for achieving integration, and aim to apply these in the context of the cultural heritage domain. The problem of converting the original data sets to RDF/OWL has not been investigated much. In this thesis we focus mostly on conversion of vocabularies. Our problem statement is as follows: How can existing vocabularies be made available to Semantic Web applications? Problems that need to be solved include understanding the original syntactical format in which the vocabulary is expressed, understanding the conceptual model that lies behind it, linking this conceptual model to that of RDF/OWL, and finding an appropriate way to convert the former model into the latter. These tasks are far from being automated, but the demand for proper conversions will increase in the coming years. Therefore, this thesis has focused on developing methods for conversion of vocabularies. Methods are step-wise processes with guidelines that can be followed by people performing the conversion task. Several choices have to be made during the process that affect the resulting representation. Conversion can be performed for the benefit of one particular application and tuned to its specific needs, but conversions can also aim at a representation that is as complete and reusable as possible for any application. The main contribution of this thesis is the development of two separate methods to cater to both situations.

In Chapter 2 we develop a first version of a generic method for conversion. The assumption is that a faithful and complete conversion of the vocabularies results in a representation that is useful for most applications. A method consisting of several steps and guidelines was drafted, and then applied to two case studies: conversions of the MeSH and WordNet vocabularies. These helped to improve the method; they showed which additional guidelines were needed to adequately handle these cases. We deliberately chose two complex vocabularies so that a broad range of vocabulary features were covered. Two tailor-made schemas for each vocabulary and a complete conversion of their content to these schemas are the outcome.

Another way to convert vocabularies suitable for a wide range of applications is to use a standard, widely supported vocabulary schema. In Chapter 3 we developed a method aimed at the emerging SKOS standard. We chose three vocabularies as use cases: a simple one (GTAA), an intermediately complex one (IPSV) and a complex case (MeSH). We found that SKOS was suitable for converting GTAA and IPSV (making use of RDF/OWL abilities to specialize a schema), but MeSH could not be covered completely because SKOS did not allow a concept’s terms to be represented as instances themselves.

In Chapter 4 we returned to the problem of generic conversion as approached in Chapter 2. The assumption that our method results in vocabularies useful for many applications was tested by comparing our generic conversion of WordNet to a conversion developed with application use cases in mind (in the W3C Semantic Web Best Practices Working Group). The comparison of the two WordNets showed how our generic method should be changed to cater for a wider range
of applications. However, another outcome of the chapter is that a generic method cannot cater
to all requirements an application might have, because it may require that content is left out or
structured differently than in the original source. We also improved the WordNet conversion to
SKOS by applying a newly developed extension that allows terms to be represented as instances
themselves. This solves most of the problems noted in the MeSH conversion to SKOS.

Given the results from Chapter 4, we developed a new method that can be used to cater to
specific applications in Chapter 5. We adapted the generic method by introducing specific steps
to determine the requirements and use cases that need to be covered. The principle of complete
and faithful conversion of the original source was dropped. The case study is the MultimediaN
E-Culture search and browsing application, for which the AAT, TGN and ULAN were converted.
This case study pointed out that the conversion made by the E-Culture team without our method
missed several pieces of information needed by the application use cases.

In Chapter 6 we continued to investigate conversions targeted at specific applications. In
this chapter we concentrated on alignment applications, which take two or more vocabularies
in RDF/OWL as input and produce mapping relations between concepts of the vocabularies. Our
analysis showed that these applications cannot handle representations as produced by our methods.
We provided a conversion technique to mitigate this problem. The analysis is part of a study on
new evaluation techniques for alignments of vocabularies. Alignment is a central ingredient of
integration as promoted by the Semantic Web community. The outcome of the study is that our
proposed techniques are better tuned to evaluating the quality of an alignment for a particular
application than existing techniques.

Integration of collections relies on integration of both vocabularies and metadata element sets.
In Chapter 7 we study how an existing metadata element set can be represented in RDF/OWL in a
way that is interoperable with vocabulary representations as advocated in this thesis. The metadata
element set is called VRA and caters specifically to cultural heritage. We show how VRA can be
implemented as a specialization of the more generic Dublin Core element set. Linking VRA
with Dublin Core allows integration of collections from different domains (television archives,
libraries, etcetera) into one search system. We also show how VRA can be specialized to reflect
that e.g. the Rijksmuseum uses e.g. ULAN as range of the “creator” element (we term this feature
collection-specific value ranges).

In summary, this thesis contributes to the integration of metadata collections in three ways.
Firstly and chiefly through the development of conversion methods for vocabularies and through
contributing actual conversions made with the methods. Secondly, through investigating how
metadata schemas can be represented in a way that allows using them together with vocabulary
representations produced by the methods. Thirdly, by contributing a study on how alignments can
be evaluated on their usefulness for particular applications.