Athanasios Velios*

Beyond databases: Linked open data for bookbinding descriptions

Abstract

Historical methodologies rely on evidence in order to support hypotheses. Bookbindings can play a key role in supporting hypotheses because they offer a plethora of physical evidence of materials and techniques. Publishing this evidence online may showcase a specific collection but it does not make the records re-usable by researchers without time-consuming work. In this paper I am looking into recent developments of the semantic web technologies which refer to re-usable data as "Linked Open Data" and examine how it is possible to employ these technologies in historical bookbinding research.

1 Introduction

Many recent discussions in the field of historical bookbinding are about newly-published online bookbinding collections. The use of the internet as a way of publishing research data has been widely adopted in many fields. The sciences (especially medicine and biology) have pioneered the various methods of sharing data online. While terminology such as Deep Data, Big Data and Web 2.0 is commonplace in the sciences, in the traditional humanities fields these terms are rarely mentioned (apart from researchers in the digital humanities). In bookbinding history these terms are almost irrelevant. This paper aims to shift the focus from the publications of individual bookbinding collections to the sharing of bookbinding data under a common web framework. It looks at the database as a vehicle for contributing reusable data to the

* ## 

bookbinding community and not as a standalone unit showcasing the collection of an institution.

2 Current bookbinding resources

It is appropriate to start this article by reviewing the scope and the features of online bookbinding databases. One of the outputs of a recent project at the Folger Shakespeare Library was the compilation of a list of online bookbinding databases and related resources.[1] This list is a good start for such a review while further resources can be found through online search engines. I approached these resources from the point of view of a researcher who is looking to retrieve information and evidence on historic bookbindings while at the same time being informed about the capacity of machine searching. Some significant points on the use of these resources are made in the following paragraphs.

2.1 Terminology

A number of these resources include textual descriptions of bindings. The existence of text, alongside (or independent of) the photographs, allows keyword searching in a resource. In these resources data has been identified and recorded by a researcher while the resource was being created. This data is offered as a means of retrieving records by other researchers. Typically this is done in a survey of the collection to accompany a digitisation project. However, the problem with keyword searching is that a spoken language word which is used as a keyword, may correspond to more than one concept (e.g. in English, *tooling* can mean both the activity of creating decoration on a book and the actual impression of a tool, or *textblock* can mean both the area of a page where the printed text is located and the sum of the leaves in a bound volume). In this case a mixture of results could be returned only part of

1 Folger Shakespeare Library, 2013

which matches the implied concept of the keyword. Another problem with keyword searching is that two spoken language words may correspond to the same concept (e.g. *endleaves* and *endpapers*, or *endbands* and *headbands* are often used interchangeably). The result is that the researcher has to repeat the keyword search multiple times, with different keywords, to ensure that a significant number of records is not missed. However, even having done this the researcher cannot be certain of the completeness of the results, because there is always a chance that some of them have been described with different keywords than the ones used in the multiple searches. The problem becomes more complex when multiple languages are involved as the same concept may have been described with a variety of words in the various languages. To reduce the impact of this problem the examined resources have been produced rigorously with consistent terminology. However, with a few exceptions, there has been no effort to separate the concepts being described from the spoken language terms to avoid the problem altogether. Although the rigorous work may offer confidence that keyword searching returns representative results within a resource, the terminology problem remains when considering resources comparatively. When consistent records of a resource do not agree with consistent records of another resource, then the value of being consistent is diminished. This is further discussed in the following paragraphs.

2.2 Cross-searching

Most of the resources examined are designed to retrieve results from a specific collection. The obvious problem here is that a researcher who needs to retrieve results from a range of collections has to repeat the search in multiple online locations. This is a very time-consuming process which concludes with the researcher manually synthesizing the results to produce a unified list of data.

2.3 Researcher generated data

When researchers examine resources, depending on the extent of the existing records, they may generate new data. This is often the case with resources which only host photographs, as any new observations made on the photographs are recorded by the researcher. It is also often the case with results synthesized after multiple searches on different resources. Typically, this data synthesis is done on a personal computer. The examined resources do not have provisions which would allow researchers to contribute the synthesized results back to the resources for others to use. A common problem in the field of bookbinding history is that researchers need to repeat work that was done in the past by other researchers because access to earlier data is impossible. This slows the development of knowledge and delays the establishment of bookbindings as a reliable source of historical evidence.

2.4 Problem specification

The points made in the previous paragraphs: a) the lack of a formal terminology in bookbinding records and b) the inability to search across resources, are two major limitations which delay research into historical bookbinding. Researchers require representative samples of data to make good conclusions about the history of bookbinding in general as well as the history of specific books. Without being able to search across resources, samples are limited to specific collections and therefore are not representative. Researchers are able to visit and search records from multiple collections, but this is so time-consuming that it reduces the output of researchers significantly. The lack of a formal terminology in these resources multiplies the amount of work researchers have to do to harmonise their sample thus further reducing the number of conclusions researchers can make. The issue of contributed data is also an important one as most of the conclusions researchers reach are explained in publications which rarely include the complete sample that a researcher has collected. When they do, the format (typically

print) is not suitable for the sample to be re-used without further time-consuming tasks and therefore it is of little use to other researchers.

In this paper I am proposing a way of publishing bookbinding-related data so that digesting existing data and contributing new data can be done more efficiently, which will lead to faster research conclusions and faster adoption of bookbinding as a key source of historical evidence. I am borrowing ideas from recent developments in the semantic web as well as ideas which matured in the fields of knowledge representation and reasoning. The next section introduces these ideas and their relevance to historical bookbinding.

3 Previous work

Cross-disciplinary research has been encouraged for many years. In this section I will try to highlight work that has been taking place in other fields of research and how that work can be of benefit to the study of historical bookbinding. The fields I am looking to borrow methodologies from are related to knowledge organisation, philosophy and computer science. I need to make clear that I have no background in any of these fields so the following sections should not be read as an introductory textbook, as they may contain out of date terminology and they will certainly exclude many important concepts in these fields which I am not using in this document. Instead the following sections should be read as a set of references about useful ideas for the study of historic bookbinding. Some useful books for following up these references are by Brachman and Levesque and Allemang and Hendler.[2] Essential and pioneering work in the field has been published by Doerr et al.[3]

2 Brachman/Levesque (2004) and Allemang/Hendler (2009)

3 Doerr et al. (2011)

3.1 Knowledge representation

Knowledge can be defined as a belief of a proposition. For example, a researcher can make this statement: "I know that Sinai manuscript Greek 277 has a raised endband". The important point here is that this statement is an abstract representation of a fact: a proposition which describes the physical object after an observation. Knowledge can be more generic. For example, a researcher can make this statement: "I know that all books written in Greek during the Byzantine period have raised endbands". This statement expresses a relation between the researcher (the person who carries the knowledge) and the proposition (the fact that all Byzantine Greek books have raised endbands). This statement shows us that the researcher believes the proposition to be true. Another researcher may well make a different statement: "I know that *not* all books written in Greek during the Byzantine period have raised endbands", which shows that the second researcher believes that the first statement is false. There is an important point to make here in that propositions (knowledge) are related to the person who believes it to be true or false. Other statements can include expressions of intention and attitude (e.g. "I hope that all books written in Greek during the Byzantine period have raised endbands" or "I like all books written in Greek during the Byzantine period"), but for historical research based on evidence, it is only necessary to focus on statements about belief.

Representation separates an idea from the way it is communicated. As discussed earlier in this document () the concept of an endband exists regardless of what words are used to communicate it. The word *endband* and the word *headband* are the so-called *symbols* of the concept endband. In applications which are relevant to historical bookbinding, thesauri of words provide the symbols for knowledge in the field. These are merely labels for the underlying concepts.

By combining the above descriptions we can define *knowledge representation* (KR) as a field of research where propositions made by people are expressed using symbols. A key purpose of KR is to allow new propositions (new knowledge) to develop based on the combination of existing knowledge. This is called *inference*, and it is closely linked to *reasoning* as discussed next.

3.2 Reasoning

In the previous example, we have this statement which we believe is true: "I know that all books written in Greek during the Byzantine period have raised endbands". Let us also suppose that we make these statements: "I know that Sinai manuscript Greek 418 was written in Greek" and "I know that Sinai manuscript Greek 418 was bound during the Byzantine period". From these statements I can make this statement "I know that Sinai manuscript Greek 418 has raised endbands". This is a simple example of *inference*. We had three propositions which were then combined to produce a new piece of knowledge without requiring any new observations.

The process of producing new knowledge as described in this example appears to be a result of a human brain processing existing knowledge. However, the new knowledge could have been produced by a machine (as noted, no new observation by a human is required) because of the use of symbols. *Reasoning* is the area of research where the formulas for automatically producing new knowledge are investigated. One way of explaining that is by drawing an analogy with arithmetic. Numbers are symbols in the same way propositions or concepts can be symbols. Numbers can be processed through formulas to produce results and similarly propositions can be processed to produce new propositions. Given a set of related propositions, new propositions can be calculated automatically according to some formulas or rules.

Propositions need to be well-defined in order for a machine to be able to process them as described in the previous point. This means that propositions need to be represented with discreet symbols in a machine. In the previous examples our symbols are spoken language words which correspond to concepts as described in a concept thesaurus. In other cases these symbols may be different. Later in this document (section 3.3), web addresses will be used as symbols instead. Simplicity of a proposition is also important, so it is generally preferable to break down complex propositions into smaller, simpler ones. In the previous example, the statement "I know that Sinai manuscript Greek 418 was written in Greek and bound during Byzantine period" was broken down to "I know that Sinai manuscript Greek 418 was written

in Greek" and "I know that Sinai manuscript Greek 418 was bound during the Byzantine period". Both these statements follow the structure of a triple:

subject	property	value
Sinai manuscript Greek 418	was written in	Greek
Sinai manuscript Greek 418	was bound during	Byzantine period

Tab. 1 Triple structure of machine friendly statements

The triple of terms *subject, property, value* are also met as *entity, attribute, value* elsewhere in the bibliography [for example, Dufton and Fenwick (2012)] and will be called *subject, predicate, object* later in this document in the specific context of RDF, the technology which is designed to make the semantic web a reality.

In the previous example, I explained that the statement: "I know that all books bound in Greece during the Byzantine period have raised endbands" was disputed by other researchers. After all, this is *my* knowledge, a proposition that *I* believe to be true; while other researchers may not believe this proposition. In the following paragraphs I am describing how this dispute can be modelled in a process so that it can lead to propositions which better reflect reality and therefore enhance scholarly knowledge.

Induction

The first researcher looked at a number of books from a collection and observed the raised endband on each one. While surveying the books, the researcher made statements such as "I know that this manuscript has a raised endband" and "I know that this manuscript was bound during the Byzantine period in Greece" for each book. When all the books in the collection were observed, the researcher realised that these two statements were made for every book without exception. Therefore, the researcher reasonably suggested that if we look at any other book written in Greek and bound during the Byzantine period, we will find a raised endband. In the field of reasoning, the process of drawing a wider conclusion from a subset of observations is called *induction*. Induction is only an indication of high probability and not

an absolute rule. In the previous example there is no guarantee that every book written in Greek and bound during the Byzantine period will have raised endbands. Induced conclusions are based on the assumption that the propositions are made on a representative sample. If, for example, the collection of books observed was an extraordinary set bound exceptionally by a workshop to have raised endbands, then the sample would not be representative and therefore the induced rule would be wrong. The second researcher looked at a different collection and made other observations: not all books written in Greek and bound during the Byzantine period had raised endbands.

Much of the research done in historic bookbinding is based on induction. Researchers typically look at small collections of books and attempt to make generalisations about history. In many cases they rely on existing inductions by other researchers, simply because of lack of time (for example, Szirmai highlights the impossibility of a representative sample in historical bookbinding research[4]). For this reason, it is important to emphasise that a representative sample is essential for the induced conclusions to be of value, as this fact is often neglected due to logistical limitations. There is no intention here to dismiss the value of research in the field so far. On the contrary, in the following paragraphs the induced conclusions from current limited samples become the key for the ongoing development of the field.

Deduction

Deduction can be considered as the opposite of *induction*. While during induction researchers make general propositions about bookbinding history from specific and partial observations on bindings, during deduction researchers use generally accepted propositions about bookbinding history to make new propositions about specific bookbindings. For example, the proposition that paper was not used as a bookblock material before the 8th century may be accepted as a rule by some researchers. When discussing a 7th century book, one could simply assume that the bookblock material will not be paper. A generally accepted rule is used to produce a new proposition (i.e. new knowl-

4 Szirmai (1999)

edge). The key point here is that the general proposition is believed by the researcher as a universal truth. We should emphasise, however, that the general proposition developed when a series of propositions (all 7th century books thus far examined do not consist of paper bookblocks) led to a reasonable expectation of what the bookblock material of early codices could be (i.e. not paper).

The researcher is confident that the 7th century bookblock cannot be made of paper and makes a new proposition: "I know that the bookblock (of this 7th century manuscript) is not made of paper". Another researcher has the chance to examine the book and to everybody's surprise, discovers that the bookblock is made of paper. The second researcher makes a new proposition: "I know that the bookblock (of this 7th century manuscript) is made of paper". This is the point were the conflict occurs. A machine interpreting the symbols of these two propositions would identify the opposing views:

subject	property	value
bookblock	is made of	not paper (other material)
bookblock	is made of	paper

Tab. 2 Two conflicting statements identified by a machine.

Clearly one of these propositions is closer to the truth than the other. Because we trust the good will and expertise of scholars and the common target of progressing knowledge, we accept that the proposition made after the direct observation of the bookblock is more likely to be true, than the proposition made after accepting an older proposition as true. Doerr et al.[5] formalised that way of prioritising propositions and placed observation *(factual argumentation)* at a higher level than belief adoption in their proposal for developing machine-friendly arguments in archaeology *(argumentation model)*.

The above conflict can lead to a revision of the accepted rule about early use of paper by shifting the cut-off century from 8th to 7th. This is a typical example of the interaction between induction – deduction and

5 Doerr et al. (2011)

an essential part of research in historical bookbinding. In the following paragraphs, I am describing how this process can be automated where the deduction rules can be extracted from a large sample of observations stored in various databases around the world.

3.3 Resource Description Framework and Linked Open Data

The Resource Description Framework (RDF) (hay, 2004) claims to propose the simplest way of storing knowledge: triples. An easy way to illustrate this claim is by attempting to reduce RDF's triples to something simpler like pairs. The statement "MS Sinai 418 bookblock" – "is made of" – "paper" is clear because it associates "paper" with "MS Sinai 418 bookblock". Any attempt to reduce this further is impossible: a) "MS Sinai 418 bookblock" – "paper", b) "MS Sinai 418 bookblock" – "is made of", c) "is made of" – "paper", are three pairs of data which by themselves do not make sense because they do not complete any associations. If it is reasonable to accept that the simplest form for communicating our data is triples, it is also reasonable to suggest RDF as the common platform for sharing data of bookbinding descriptions. RDF has been proposed by the World Wide Web Consortium as a web technology and it is designed for publishing data online.

It is critical to emphasise that adopting RDF for sharing data does not require any internal restructuring of existing databases. Existing records stored in databases can be exported and published as RDF. This involves some development work but it does not break any existing data-model which large institutions may depend on and therefore may be reluctant to modify.

Another important idea in RDF is the *identification* of associated things. In the statement "MS Sinai 418 bookblock" – "is made of" – "paper", the association is between "paper" and "MS Sinai 418 bookblock" through the property (also called the *predicate*) "is made of". As explained earlier these are symbols of underlying concepts in a specific spoken language. Because a machine does not comprehend the meaning of these symbols (e.g. "paper" as in organic material instead of "paper" as in newspaper), it would be useful if we could separate the two

by giving them a unique identifier which can be used by a machine. Because the RDF technology has been widely used in the web, an existing method of identifying things online can be used to point to the right concept. This method is the system of web address that we use widely, a *Uniform Resource Identifier* or a *URI* (for information about URIs, URLs and URNs see URI Planning Interest Group 2001, for this paper there is no need to discuss their differences). For example Ligatus's website holds a record of "MS Sinai 418 bookblock" at the address:

www.ligatus.org.uk/stcatherines/ms/0418/bookblock

We could also use another reference for "paper" material such as the one from Wikipedia (in fact it is from dbpedia, an RDF-friendly version of Wikipedia):

http://dbpedia.org/page/Paper

Let us now transform the original statement by replacing the English language symbols with the symbols from the web addresses:

subject – www.ligatus.org.uk/stcatherines/ms/0418/bookblock
predicate – is made of
object – http://dbpedia.org/page/Paper

Clearly we have yet to replace the predicate "is made of" with another web address reference, but the above statement is much clearer to a machine because, although the machine still does not understand what we are talking about, it can at least process the symbols and point us in the right direction. There is little value in going to all this trouble for a single statement. The true value of RDF is evident when we have millions of these statements and we can then search through them to identify all the books with a bookblock made of paper and, of course refine our query with further criteria. The important thing here is that we can perform this search on many databases across various institutions provided that they have published their data in RDF, and I em-

phasise again that this does not require changing the actual databases (also see).

Publishing book descriptions as sets of RDF statements allows researchers to make more accurate propositions based on induction because they have a larger sample of data available. Saved queries providing evidence for these propositions can be tested regularly and revised automatically as new data become available and therefore the identification and provenance of books will become more accurate.

The term *Linked Open Data* describes data that is published as RDF online and is available for other data to link to (in the same way the Sinai bookblock is linked to the Wikipedia article on paper material). The point of availability relies on the data being made available without copyright restrictions which will be discussed in the section. For further introductory reading on RDF see **man, 2004**

4 Standard resources

In the previous section I highlighted ideas and technologies which have the potential to change the nature of research in historical bookbinding. The examples mentioned were illustrative to explain how the technology works. In this section I will refer to specific online resources and projects and their capacity to be used for bookbinding resources expressed in RDF.

4.1 Bookbinding concepts

Much of the description of bookbindings involves identifying discreet pieces of material on a binding or discreet modifications of materials and assigning a *type* (or *class*) to them. A piece of material placed against the back of the bookblock to strengthen the binding structure is associated with type “sewing support”. The modification of the material of a fastening strap in a typical Byzantine binding is associated with type “braided”. As mentioned earlier we need URIs (or web addresses) to be able to refer to these types. Dbpedia is certainly an option, how-

ever researchers may find other specialised resources more suitable. A popular option is the dictionary by Roberts et al.[6] whose electronic version makes sure that each term of the resource corresponds to a unique web address and as such it can be used as a reference. Roberts et al. have produced a dictionary of English terms rather than a concept thesaurus, nevertheless much of the terminology included in databases, especially in English speaking languages is probably covered by it, although care is required to ensure that database records match the term definitions as far as meaning is concerned. Another important resource in the field is the Arts and Architecture Thesaurus (AAT, The Getty Research The Getty Research Institute) which includes a number of bookbinding terms. At the time of writing this document, AAT is not available in a suitable format for RDF referencing, however the Getty Research Institute has laid a plan for making such linking available[7] which may well be complete by the time this document is published. A recent European network of bookbinding experts[8] has come together to produce a bookbinding thesaurus for RDF data publishing. At the time of writing this document, this work is progressing and nearing completion, and it may well be that by the time this document is published, the LoB thesaurus will have been published.

Therefore when an existing resource is to be published as Linked Open Data, the first step is to ensure that the details stored on the database match the meaning of the concepts/terms of a suitable thesaurus.

4.2 Ontological framework

In section 3.3 Resource Description Framework and Linked Open Data, I discussed replacing the subject and object of each triple with a web address which is a unique reference to the book component or concept used. In this section I discuss the replacement of predicates. In the ex-

6 Roberts et al. (1982)

7 The Getty Research Institute (2013)

8 Language of Bindings (LoB) network, led by Ligatus at the University of the Arts London and funded by the Arts and Humanities Research Council in the UK.

ample from section 3.3 the predicate is "is made of". This describes a typical question researchers ask about books, i.e. what material is it made of. Many other predicates exist, such as "is owned by", i.e. which is the organisation or person who currently has custody of the book. Or "was created by", i.e. what is the origin of the book. Many of these questions are asked for historical objects in general, not only for books. A long established project by the International Council of Museums (ICOM)[9] and its International Committee for Documentation (CIDOC)[10] is the Conceptual Reference Model (CRM) as described in cro (2011). The CIDOC-CRM is a significant specification which has been adopted by many organisations in the world and has also become an ISO standard. The maturity of the project ensures that many of the requirements researchers have for expressing their propositions are covered and therefore it is a good ontological framework for book description. The CIDOC-CRM specifies a set of entities which can be used in RDF triples and can replace the predicates in every statement (as well as the subject and object if required). The predicate "is made of" is formalised in CIDOC-CRM with the entity "P45 consists of (is incorporated in)". "P45" is an internal identifier for the specific entity which is often used as a short name for it. "consists of" is the English label of the entity and it is useful to make sense of a statement to humans, for example:

subject – www.ligatus.org.uk/stcatherines/ms/0418/bookblock
predicate – P45 consists of
object – http://dbpedia.org/page/Paper

The additional statement "(is incorporated in)" allows the statement to be read in both directions, which is unusual, such as:

subject – http://dbpedia.org/page/Paper
predicate – P45 is incorporated in
object – www.ligatus.org.uk/stcatherines/ms/0418/bookblock

9 http://icom.museum/

10 http://network.icom.museum/cidoc/

It is important to emphasise that the CIDOC-CRM specifies entities and their reasonable relations. It defines a set of rules which match the real world as perceived by museum professionals. For example a book-block (E18 Physical Thing) can have a material (E57 Material) through the property "P45 consists of". The text of a marginal note on the book-block (E34 Inscription) cannot have a material because it is an abstract concept. Instead a statement can link the bookblock (E18 Physical Thing and E84 Information carrier) to the note (E34 Inscription) with the property "P128 carries". The list of RDF statements which could be used to communicate these propositions follow:

Specify the bookblock is a Physical Thing as per the CRM's definitions:

subject – www.ligatus.org.uk/stcatherines/ms/0418/bookblock
predicate – www.w3.org/1999/02/22-rdf-syntax-ns#type
object – www.cidoc-crm.org/rdfs/cidoc_crm_v5.0.2_english_label.rdfs#E18.Physical_Thing

Specify that the bookblock is made of paper:

subject – www.ligatus.org.uk/stcatherines/ms/0418/bookblock
predicate – www.cidoc-crm.org/rdfs/cidoc_crm_v5.0.2_english_label.rdfs#P45.consists_of
object – http://dbpedia.org/page/Paper

Specify that the bookblock is also an Information Carrier, as per the CRM's definitions:

subject – www.ligatus.org.uk/stcatherines/ms/0418/bookblock
predicate – www.w3.org/1999/02/22-rdf-syntax-ns#type
object – www.cidoc-crm.org/rdfs/cidoc_crm_v5.0.2_english_label.rdfs#E84.Information_Carrier

Specify that the marginal note is an Inscription as per the CRM's definitions:

subject – www.ligatus.org.uk/stcatherines/ms/0418/bookblock/marginal_note_1
predicate – www.w3.org/1999/02/22-rdf-syntax-ns#type
object – www.cidoc-crm.org/rdfs/cidoc_crm_v5.0.2_english_label.rdfs#E34.Inscription

Specify that the marginal note is carried by the bookblock:

subject – www.ligatus.org.uk/stcatherines/ms/0418/bookblock
predicate – www.cidoc-crm.org/rdfs/cidoc_crm_v5.0.2_english_label.rdfs#P128.carries
object – www.ligatus.org.uk/stcatherines/ms/0418/bookblock/marginal_note_1

This list of statements may look complicated for a rather simple description. However, a machine reading these statements is unable to comprehend the context and much of the information that is implied during research has to be made explicit. The result is a large number of statements which although may be time-consuming for humans to read, can be processed quickly by a machine. This large number of statements is the basis for induction where researchers are able to query large volumes of data to produce general conclusions and therefore RDF is an interesting tool which could transform research in bookbinding history.

5 Discussion

The previous sections described the process of reasoning through induction/deduction in bookbinding research and highlighted the problems of undertaking that process manually, i.e. by examining a small number of records, relying on opinions of other people without being able to confirm their resources and therefore risking the accuracy of new knowledge. RDF is a technology which is transforming the way

propositions can be made and allows researchers to query large number of records as well as share their own conclusions in a standard way. Although RDF is certainly a technology which bookbinding experts and related institutions should adopt for their bookbinding descriptions, there are some unresolved issues which require discussion.

5.1 Schemas

In section 3.3 Resource Description Framework and Linked Open Data, I stated that publishing data held at existing databases as RDF will not break any compatibility with institutional models and structures. On the other hand, as mentioned in much of the information in existing databases is implied. For example a description of a binding as "Sewn on three recessed hemp cord supports that do not correspond to the five false bands on the spine" implies several pieces of data: a) there is sewing thread, b) supports are made of hemp cord, c) supports are recessed, d) binding has five false bands, etc. This information is understood by a bookbinding expert reading the document but would not be useful in a machine search unless the information can be broken down as separate statements. Certainly, this record can be published as RDF but only as a note of a description and therefore it cannot contribute to inferencing. Once the value of RDF is appreciated, it would lead to database structures (also called *schemas*) being revised to offer more data. Consequently the decision to publish data as RDF could trigger larger projects in institutions and the improvement of database structures.

5.2 Copyright

As mentioned before, Linked Open Data require free access. The issue of copyright for Linked Open Data has attracted attention because if institutions or individuals want to prevent people from using their data, Linked Open Data does not offer any suitable mechanism. The publishers of Linked Open Data can require the use of a license by their users, it is however extremely difficult to enforce it and almost

always that license offers the resource royalty-free. In some respects a strategic decision by an institution to publish its data as Linked Open Data implies that the use of the institution's resources is encouraged and therefore issues of copyright are almost irrelevant.

5.3 Provenance

A current discussion in RDF is the documentation of the *provenance* of RDF statements. Provenance in the context of RDF includes the information needed to establish the quality and authenticity of the statement. At the moment anybody can publish RDF statements and apart from the fact that they need to be published through an online resource (e.g. a SPARQL endpoint[11]) it is impossible to establish who made the statement and whether it is authentic. For resources published by institutions, researchers may be re-assured for quality. However, in academic work clearer information about the owner of each statement and the date of the statement could be useful when researchers begin publishing their data as RDF regularly. The W3C has not made any recommendations on how this could be done, but a strong proposal suggests the addition of an extra label to an RDF triple to identify it (essentially making it a quad). Further RDF statements can then be made about that quad which will capture provenance data. However, the authenticity of a statement will still remain an issue. Among others, digital signatures have been suggested as ways of establishing authenticity. For further reading on this issue see Carroll et al.[12] At the moment there is some concern surrounding the use of RDF resources due to a lack of provenance data. However, given that most datasets are published by established institutions it is reasonable to accept them as reliable.

11 See W3C SPARQL Working Group (2013)

12 Carroll et al. (2005)

5.4 Searching

This document focused on structuring bookbinding records to allow more accurate conclusions about historic bookbinding. It did not focus on how it is possible to search RDF data. An earlier reference to SPARQL[13] indicated that this is the means to query RDF triples. Although it is expected that user-friendly tools for searching will gradually become available, at the moment a basic knowledge of SPARQL is required for searching published resources.

6 Closing remarks

I hope that with this document I have illustrated the relevance of new Semantic Web technologies, such as RDF, to research in historic bookbinding. This is an exciting development which is worth serious consideration as more and more records of bindings are made available online. The existence of critical resources, such as the Getty AAT and the CIDOC-CRM are extremely helpful in the field of bookbinding history, because they have already delivered much of the foundation work. Publishing bookbinding descriptions and conclusions about the history of the book as Linked Open Data will bring the disparate databases of the different institutions together and offer opportunities for online research which have never before been available to bookbinding historians. Such opportunities will allow the field to make rapid progress in its findings.

Acknowledgments

I would like to thank Martin Doerr of the Foundation of Research and Technology Hellas for his advice on the semantic technologies discussed and the thinking behind the philosophical issues described. I would also like to thank Heather Ravenberg for her work on the CI-

13 W3C SPARQL Working Group (2013)

DOC-CRM and Prof. Nicholas Pickwoad for sharing his expertise on bookbinding. This work was supported by the Ligatus research centre at the University of the Arts London.

References

RDF semantics, February 2004 – www.w3.org/TR/2004/REC-rdf-mt-20040210/.

RDF primer, February 2004. – www.w3.org/TR/2004/REC-rdf-primer-20040210/.

Definition of the CIDOC conceptual reference model, December 2011.

Allemang/Hendler (2009) Dean Allemang and James Hendler: Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL. Morgan Kaufmann, April 2009

Brachmann/Levesque (2004) Ronald Brachman and Hector Levesque: Knowledge Representation and Reasoning. Elsevier, June 2004

Carroll et al. (2005) Jeremy J. Carroll, Christian Bizer, Pat Hayes, and Patrick Stickler: Named graphs, provenance and trust. In Proceedings of the 14th international conference on World Wide Web, WWW '05, pp. 613–622, New York, NY, USA, 2005 – http://doi.acm.org/10.1145/1060745.1060835.

Doerr et al. (2011) Martin Doerr, Athina Kritsotaki, and Katerina Boutsika. Factual argumentation – a core model for assertions making. J. Comput. Cult. Herit., 3 (3): 8:1–8:34, February 2011. – http://doi.acm.org/10.1145/1921614.1921615.

Dufton/Fenwick (2012) Andrew Dufton and Corisande Fenwick: Beyond the grave. Developing new tools for medieval cemetery analysis at Villamagna, Italy. In: Angeliki Chrysanthi, Patricia Murrieta-Flores, and Constantinos Papadopoulos (editors): Thinking Beyond the Tool. Archaeological Computing and the Inter-

pretive Process, volume 2344, pp. 155–167. Archaeopress, Oxford, March 2012

Folger Shakespeare Library (2013) Folger Shakespeare Library. Other resources related to binding images-Folger Shakespeare library, 2013. – www.folger.edu/Content/Collection/Folger-Bindings-Image-Collection/Other-Resources-Related-to-Binding-Images.cfm.

Roberts et al. (1982) Matt Roberts, Don Etherington, Margaret R. Brown: Bookbinding and the conservation of books: a dictionary of descriptive terminology. Library of Congress: For sale by Supt. of Docs., U.S. G.P.O., Washington, 1982.

Szirmai (1999) J. A. Szirmai: The archaeology of medieval bookbinding. Ashgate, Aldershot, Hants.; Brookfield, Vt., 1999.

The Getty Research Institute (2013) The Getty Research Institute. Art & architecture thesaurus. – www.getty.edu/research/tools/vocabularies/aat/index.html.

Getty Research Institute. Getty vocabularies as LOD, July 2013. – www.getty.edu/research/tools/vocabularies/lod/index.html.

URI Planning Interest Group URI Planning Interest Group. URIs, URLs, and URNs: clarifications and recommendations 1.0, September 2001. – www.w3.org/TR/uri-clarification/.

W3C SPAQL Working Group (2013) W3C SPAQL Working Group. SPARQL 1.1 overview, March 2013. – www.w3.org/TR/sparql11-overview/.