INTRODUCTION

There are a great number of SPARQL tutorials on the Web, but the majority of them make at least one of the following two assumptions- which do not always hold true in real-life use cases:

1) That the dataset the user wants to query is relatively small (e.g. "toy" examples)

2) That if the user is querying a massive database (e.g., DBpedia), a SPARQL endpoint will be provided.

What does the user do when he or she discovers that their dataset is available in an RDF format, but as a data dump – and that it contains over twenty million triples?

There are many different tools available for storing and querying RDF data, and the right one for the job depends on how the data will ultimately be used. This tutorial represents only one possible solution; its primary intention is to allow the user to retrieve a dataset and start exploring it as quickly as possible and, hopefully, in a way that is simple enough for novice users.

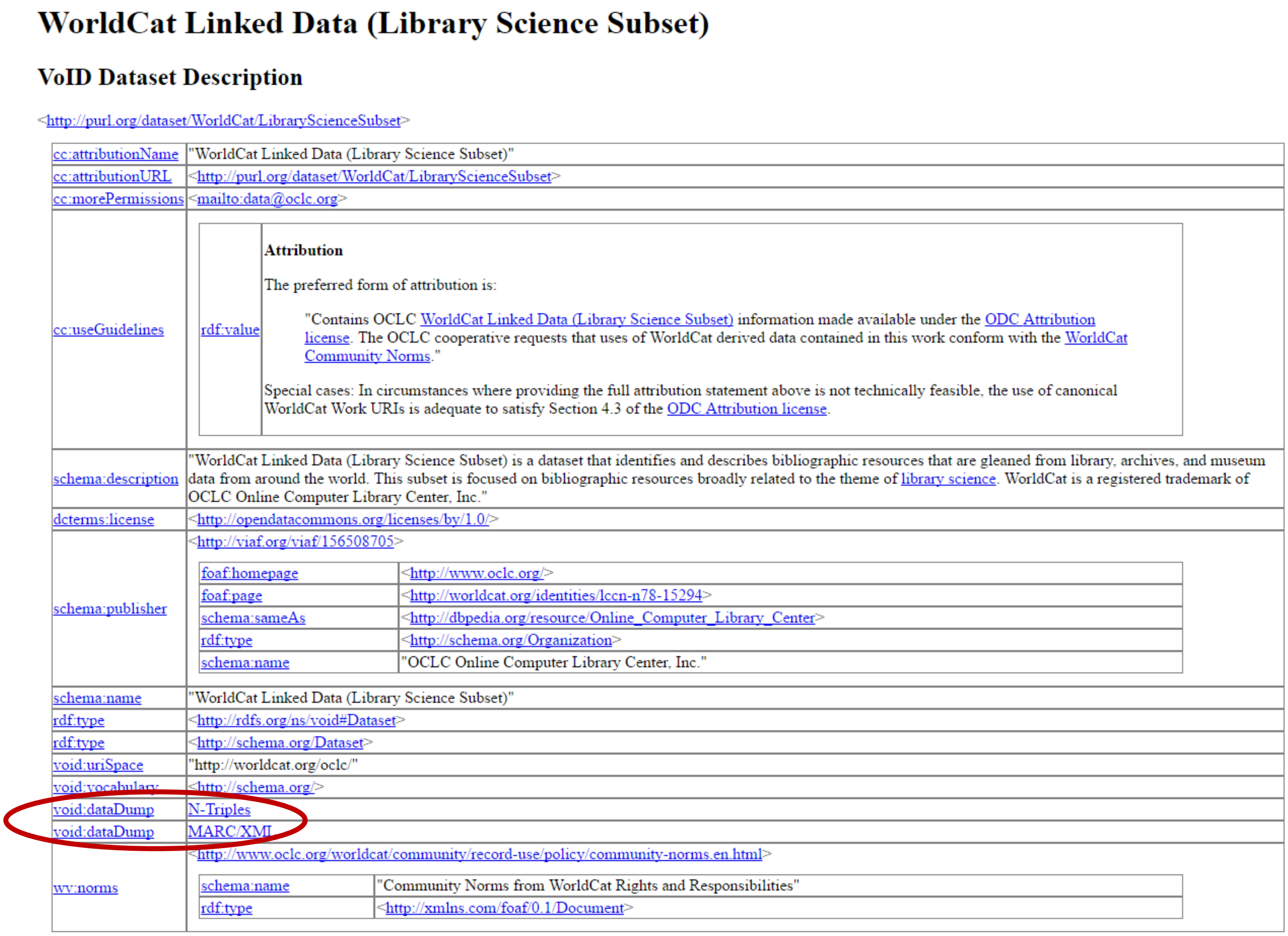
GETTING THE DATA

Let's say that a colleague has given you a link to a dataset: ftp://anonftp.oclc.org/pub/researchdata/libsci/void-WorldCatLibSci.html

Upon arriving at the webpage, you see (**Figure 1**). There is a lot of useful information here, particularly about the publisher and how the dataset may be used.

As you skim further down the VoID description, you will find the information about how to actually access the dataset. For some datasets, you might see an attribute *"void:sparqlEndpoint"* with a URL (e.g., http://dbpedia.org/sparql) pointing to the service.

Other times, you will see - either instead of or in addition to the endpoint- the attribute *"void:dataDump"*. This means that you can physically download a copy of the dataset to your machine. In this case, there are two formats available for downloading the file: "N-triples" or "MARC/XML". You will want to choose *"N-triples"*, a common RDF serialization that Jena TDB can process.



WorldCat Linked Data (Library Science Subset)

VoID Dataset Description

<http://purl.org/dataset/WorldCat/LibraryScienceSubset>

<table>
<tr><td>cc:attributionName</td><td>"WorldCat Linked Data (Library Science Subset)"</td></tr>
<tr><td>cc:attributionURL</td><td><http://purl.org/dataset/WorldCat/LibraryScienceSubset></td></tr>
<tr><td>cc:morePermissions</td><td><mailto:data@oclc.org></td></tr>
<tr><td>cc:useGuidelines</td><td><table><tr><td>rdf:value</td><td><b>Attribution</b>
The preferred form of attribution is:
"Contains OCLC WorldCat Linked Data (Library Science Subset) information made available under the ODC Attribution license. The OCLC cooperative requests that uses of WorldCat derived data contained in this work conform with the WorldCat Community Norms."
Special cases: In circumstances where providing the full attribution statement above is not technically feasible, the use of canonical WorldCat Work URIs is adequate to satisfy Section 4.3 of the ODC Attribution license.</td></tr></table></td></tr>
<tr><td>schema:description</td><td>"WorldCat Linked Data (Library Science Subset) is a dataset that identifies and describes bibliographic resources that are gleaned from library, archives, and museum data from around the world. This subset is focused on bibliographic resources broadly related to the theme of library science. WorldCat is a registered trademark of OCLC Online Computer Library Center, Inc."</td></tr>
<tr><td>dcterms:license</td><td><http://opendatacommons.org/licenses/by/1.0/></td></tr>
<tr><td>schema:publisher</td><td><http://viaf.org/viaf/156508705><table><tr><td>foaf:homepage</td><td><http://www.oclc.org/></td></tr><tr><td>foaf:page</td><td><http://worldcat.org/identities/lccn-n78-15294></td></tr><tr><td>schema:sameAs</td><td><http://dbpedia.org/resource/Online_Computer_Library_Center></td></tr><tr><td>rdf:type</td><td><http://schema.org/Organization></td></tr><tr><td>schema:name</td><td>"OCLC Online Computer Library Center, Inc."</td></tr></table></td></tr>
<tr><td>schema:name</td><td>"WorldCat Linked Data (Library Science Subset)"</td></tr>
<tr><td>rdf:type</td><td><http://rdfs.org/ns/void#Dataset></td></tr>
<tr><td>rdf:type</td><td><http://schema.org/Dataset></td></tr>
<tr><td>void:uriSpace</td><td>"http://worldcat.org/oclc/"</td></tr>
<tr><td>void:vocabulary</td><td><http://schema.org/></td></tr>
<tr><td>void:dataDump</td><td>N-Triples</td></tr>
<tr><td>void:dataDump</td><td>MARC/XML</td></tr>
<tr><td>wv:norms</td><td><http://www.oclc.org/worldcat/community/record-use/policy/community-norms.en.html><table><tr><td>schema:name</td><td>"Community Norms from WorldCat Rights and Responsibilities"</td></tr><tr><td>rdf:type</td><td><http://xmlns.com/foaf/0.1/Document></td></tr></table></td></tr>
</table>

Figure 1- VoID Description of OCLC Dataset

After downloading the file, you will notice that it has two extensions, *“.nt”* (for *“N-Triples”)* and *“.gz”* (for *“gzip”*, a type of tape archive file). To unzip this type of file, you will need to use freely available software, such as *WinZip* or *7-Zip*. You can now save the regular *“.nt”* file in a convenient directory for further processing.