Bulk Data Ingest Process for Archonnex Platform (SIP)

Archonnex support OAIS reference model and refers to data and metadata submitted by data producers as Submission Information Package (SIP). System supports projects, folders and files as it's organizational units making it easy to zip and import contents from a computer directory and its contents into the system. Root level of the zip is treated as a project and system will assign a unique project ID when imported into the system, name of the zip file will be ignored. Root level must contain a file with name "metadata.rdf" with RDF xml detailing metadata using standard+icpsr ontologies. There are very few required fields like Author (person submitting the content), Title of the Project, Principal Investigators and Summary/Abstract . This metadata file can capture metadata at all levels within SIP using RDF XML syntax. When zip file contents are exploded system will apply URL escape mechanism to replace special characters and whitespaces with hyphens in folder and file names, but original names will be retained in title metadata. File mime types and MD5 hash can be specified in metadata.rdf file for validation, if not provided system computes this information during ingestion. Unknown formats will be recorded as binaries and password protected/encrypted files may get rejected if virus scanner is unable to scan it.

Citations can be included as valid DOI URLs from where it can be imported directly into the system. But if there are citations with no valid URLs behind to support the import, they can be included in the zip using a special folder named "Citations" and individual citations in RIS format files, each citation will be an individual file with name pattern like this "xxxx.ris", these file URIs should be referred to in the metadata.rdf file as a resource.

Full RDF Mapping can be found here.

Below you can find a sample metadata.rdf file and java code that produced it.Sample zip file used for demo is attached here.

```
<rdf:RDF
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:disco="http://rdf-vocabulary.ddialliance.org/discovery#"
   xmlns:schema="http://schema.org/"
   xmlns:icpsr="http://icpsr.umich.edu/study#"
   xmlns:dcterms="http://purl.org/dc/terms/"
   xmlns:dc="http://purl.org/dc/elements/1.1/"
   xmlns:imeta="http://icpsr.umich.edu/imeta#"
   xmlns:ebucore="http://www.ebu.ch/metadata/ontologies/ebucore/ebucore#"
   xmlns:premis="http://www.loc.gov/premis/rdf/v1#"
   xmlns:citeproc="https://github.com/citation-style-language/schema/blob/master/csl-data#"
   xmlns:ore="http://www.openarchives.org/ore/terms/"
   xmlns:skos="http://www.w3.org/2004/02/skos/core#"
```

```
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xmlns:dcat="http://www.w3.org/ns/dcat#"
    xmlns:cito="http://purl.org/spar/cito#"
    xmlns:foaf="http://xmlns.com/foaf/0.1/" >
  <rdf:Description rdf:about="file://root/project/timeperiod/0">
    <schema:endDate>2018-12-31</schema:endDate>
    <schema:startDate>2015-01-01</schema:startDate>
    <dcterms:description>Between year 2015 and 2018</dcterms:description>
  </rdf:Description>
  <rdf:Description rdf:about="file://root/project/inner-folder">
    <dcterms:description>This is an inner folder with more data files</dcterms:description>
  </rdf:Description>
  <rdf:Description rdf:about="file://root/project#timeperiods">
    <rdf:_1>file://root/project/timeperiod/0</rdf:_1>
    <rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Seq"/>
  </rdf:Description>
  <rdf:Description rdf:about="file://root/project">
    <imeta:timeperiods>file://root/project#timeperiods</imeta:timeperiods>
    <imeta:fundingSources>file://root/project#fundingSources</imeta:fundingSources>
    <imeta:publications>file://root/project#publications</imeta:publications>
    <imeta:distributor>file://root/project#distributor</imeta:distributor>
    <dcterms:description>&lt;p&gt;This may contain hhtml content describing the
resources&lt;/p&gt;</dcterms:description>
    <dcterms:creator>file://root/project#creator</dcterms:creator>
    <dcterms:agent>file://root/project/people/0</dcterms:agent>
    <dcterms:alternative>This is an alternative title</dcterms:alternative>
    <dcterms:title>This is a sample title</dcterms:title>
  </rdf:Description>
  <rdf:Description rdf:about="file://root/project#distributor">
    <rdf:_2>University of State</rdf:_2>
    <rdf:_1>ICPSR</rdf:_1>
    <rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Seq"/>
  </rdf:Description>
  <rdf:Description rdf:about="file://root/project#publications">
    <rdf:_2>file://root/project/citations/citation_94551.ris</rdf:_2>
    <rdf:_1>file://root/project/citations/citation_94550.ris</rdf:_1>
    <rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Seq"/>
  </rdf:Description>
  <rdf:Description rdf:about="file://root/project#fundingSources">
    <rdf:_1>file://root/project/fundingSource/0</rdf:_1>
    <rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Seq"/>
  </rdf:Description>
  <rdf:Description rdf:about="file://root/project/people/1">
    <foaf:member>University of County</foaf:member>
    <foaf:mbox>test@email.com</foaf:mbox>
    <foaf:lastName>Spade</foaf:lastName>
    <foaf:firstName>Kate</foaf:firstName>
  </rdf:Description>
```

<rdf:Description rdf:about="file://root/project/organization/0">
<icpsr:orgName>University of Example</icpsr:orgName>
<icpsr:orgId>23233</icpsr:orgId>
</rdf:Description>
<rdf:Description rdf:about="file://root/project/people/0">
<foaf:member>University of State</foaf:member>
<foaf:mbox>test@email.com</foaf:mbox>
<foaf:lastName>Doe</foaf:lastName>
<foaf:firstName>John</foaf:firstName>
</rdf:Description>
<rdf:Description rdf:about="file://root/project#creator">
<rdf:_3>file://root/project/people/1</rdf:_3>
<rdf:_2>file://root/project/people/0</rdf:_2>
<rdf:_1>file://root/project/organizations/0</rdf:_1>
<rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Seq"/>
</rdf:Description>
<rdf:Description rdf:about="file://root/project/inner-folder/codebook.pdf">
<icpsr:MD5>5a4eb3dcf78648dbf3ab0070e6d82718</icpsr:MD5>
<icpsr:fileContentType>application/pdf</icpsr:fileContentType>
<dcterms:description>This is code book PDF document within inner-folder</dcterms:description>
</rdf:Description>
<rdf:Description rdf:about="file://root/project/fundingSource/0">
<dcterms:title>Grant Number 232323</dcterms:title>
<foaf:organization>Federal Agency NA</foaf:organization>
</rdf:Description>
</rdf:RDF>

Java Code Sample

```
package edu.umich.icpsr.test;

import com.hp.hpl.jena.rdf.model.Model;
import com.hp.hpl.jena.rdf.model.ModelFactory;
import com.hp.hpl.jena.rdf.model.Resource;
import com.hp.hpl.jena.rdf.model.Seq;

public class RDFGenerator {
	public static void main(String[] args) {
		try {
			Model model = ModelFactory.createDefaultModel();
			String dcterms = "http://purl.org/dc/terms/";
			String foaf = "http://xmlns.com/foaf/0.1/";
			String imeta = "http://icpsr.umich.edu/imeta#";
			String disco = "http://rdf-vocabulary.ddialliance.org/discovery#";
			String schema = "http://schema.org/";
			String icpsr = "http://icpsr.umich.edu/study#";
			String dc = "http://purl.org/dc/elements/1.1/";
			String ebucore = "http://www.ebu.ch/metadata/ontologies/ebucore/ebucore#";
			String premis = "http://www.loc.gov/premis/rdf/v1#";
			String citeproc = "https://github.com/citation-style-language/schema/blob/master/csl-data#";
			String skos = "http://www.w3.org/2004/02/skos/core#";
			String xsi = "http://www.w3.org/2001/XMLSchema-instance";
			String ore = "http://www.openarchives.org/ore/terms/";
			String dcat = "http://www.w3.org/ns/dcat#";
			String cito = "http://purl.org/spar/cito#";

			model.setNsPrefix("dcterms", dcterms);
			model.setNsPrefix("dc", dc);
			model.setNsPrefix("disco", disco);
			model.setNsPrefix("schema", schema);
			model.setNsPrefix("imeta", imeta);
			model.setNsPrefix("icpsr", icpsr);
			model.setNsPrefix("ebucore", ebucore);
			model.setNsPrefix("premis", premis);
			model.setNsPrefix("citeproc", citeproc);
			model.setNsPrefix("skos", skos);
			model.setNsPrefix("xsi", xsi);
			model.setNsPrefix("ore", ore);
			model.setNsPrefix("dcat", dcat);
			model.setNsPrefix("cito", cito);
			model.setNsPrefix("foaf", foaf);

			String projectUri = "file://root/project";
			Seq distributors = model.createSeq(projectUri + "#distributor");
			distributors.add("ICPSR");
			distributors.add("University of State");
			Seq pis = model.createSeq(projectUri + "#creator");
			pis.add(projectUri + "/organizations/0");
			pis.add(projectUri + "/people/0");
			pis.add(projectUri + "/people/1");
			Resource pi1 = model.createResource(projectUri + "/people/0");
			pi1.addProperty(model.createProperty(foaf, "firstName"), "John");
			pi1.addProperty(model.createProperty(foaf, "lastName"), "Doe");
			pi1.addProperty(model.createProperty(foaf, "mbox"), "test@email.com");
			pi1.addProperty(model.createProperty(foaf, "member"), "University of State");

			Resource pi2 = model.createResource(projectUri + "/people/1");
			pi2.addProperty(model.createProperty(foaf, "firstName"), "Kate");
			pi2.addProperty(model.createProperty(foaf, "lastName"), "Spade");
			pi2.addProperty(model.createProperty(foaf, "mbox"), "test@email.com");
			pi2.addProperty(model.createProperty(foaf, "member"), "University of County");
```

```
                Resource org1 = model.createResource(projectUri + "/organization/0");
                org1.addProperty(model.createProperty(icpsr, "orgId"), "23233");
                org1.addProperty(model.createProperty(icpsr, "orgName"), "University of Example");

                Resource timePer1 = model.createResource(projectUri + "/timeperiod/0");
                timePer1.addProperty(model.createProperty(dcterms, "description"), "Between year 2015 and
2018");
                timePer1.addProperty(model.createProperty(schema, "startDate"), "2015-01-01");
                timePer1.addProperty(model.createProperty(schema, "endDate"), "2018-12-31");

                Resource fnd1 = model.createResource(projectUri + "/fundingSource/0");
                fnd1.addProperty(model.createProperty(foaf, "organization"), "Federal Agency NA");
                fnd1.addProperty(model.createProperty(dcterms, "title"), "Grant Number 232323");

                Seq pub1 = model.createSeq(projectUri + "#publications");
                pub1.add(projectUri + "/citations/citation_94550.ris");
                pub1.add(projectUri + "/citations/citation_94551.ris");

                Seq fndSources = model.createSeq(projectUri + "#fundingSources");
                fndSources.add(projectUri + "/fundingSource/0");

                Seq tmPeriods = model.createSeq(projectUri + "#timeperiods");
                tmPeriods.add(projectUri + "/timeperiod/0");

                Resource project = model.createResource(projectUri);
                project.addProperty(model.createProperty(dcterms, "title"), "This is a sample title");
                project.addProperty(model.createProperty(dcterms, "alternative"), "This is an alternative title");
                project.addProperty(model.createProperty(dcterms, "agent"), projectUri + "/people/0");
                project.addProperty(model.createProperty(dcterms, "creator"), projectUri + "#creator");
                project.addProperty(model.createProperty(dcterms, "description"), "<p>This may contain hhtml
content describing the resources</p>");
                project.addProperty(model.createProperty(imeta, "distributor"), projectUri + "#distributor");
                project.addProperty(model.createProperty(imeta, "publications"), projectUri + "#publications");
                project.addProperty(model.createProperty(imeta, "fundingSources"), projectUri +
"#fundingSources");
                project.addProperty(model.createProperty(imeta, "timeperiods"), projectUri + "#timeperiods");

                Resource folder = model.createResource(projectUri + "/inner-folder");
                folder.addProperty(model.createProperty(dcterms, "description"), "This is an inner folder with more
data files");

                Resource file = model.createResource(projectUri + "/inner-folder/codebook.pdf");
                file.addProperty(model.createProperty(dcterms, "description"), "This is code book PDF document
within inner-folder");
                file.addProperty(model.createProperty(icpsr, "fileContentType"), "application/pdf");
                file.addProperty(model.createProperty(icpsr, "MD5"), "5a4eb3dcf78648dbf3ab0070e6d82718");

                model.write(System.out);
            } catch (Exception e) {
                // TODO: handle exception
                e.printStackTrace();
            }
        }
}
```