

The challenge is to build a metadata extractor for all the project Gutenberg titles which are available here: <https://www.gutenberg.org/wiki/Gutenberg:Feeds>  
(<https://www.gutenberg.org/cache/epub/feeds/rdf-files.tar.zip>)

Each book has an RDF file which will need to be processed to extract the:

- id (will be a number with 0-5 digits)
- title
- author/s
- publisher (value will always be Gutenberg)
- publication date
- language
- subject/s
- license rights

Note: For some books all of the data won't be available.

Your tasks are:

- Write a function that reads a single file in and outputs the correct output, using something like <https://www.npmjs.com/package/xml2js> or <https://www.npmjs.com/package/xmlDOM> will probably be useful to read the rdf files
- Store the output in a database of your choice locally for later querying, perhaps something like <https://github.com/sequelize/sequelize> with MySQL/PostGres or use something else!
- Write some unit tests in mocha for the code, use <https://www.npmjs.com/package/istanbul> to measure the code coverage
- Run the function against all the rdf files
- Send through the code once you're done!

Important aspects to consider

- Scalability, how long does it take to index all the content
- Reliability, does all the information process correctly?
- Querying the dataset, how should the database be configured to search for specific fields quickly?
  - Title
  - Author name
  - Publication date