Given the set of bibliographic records exposed via the OAI-PMH protocol by the Institutional repository of the University of Bielefeld [1], write a maven project to build a Java program that

1. Collects all the ~73K XML records.
2. Defines a Spark data frame (or dataset, as the candidate prefers) that defines the following data model and populate it extracting the following information from the XML records

* record identifier, it can be extracted using the XPath //\*[local-name()='header']/\*[local-name()='identifier']
* Record typology (Journal Article, Conference object, book, etc…), the value can be extracted using the XPath: //\*[local-name()='resourceType']
* Publication year, the value can be extracted using the XPath: //\*[local-name()='date' and @dateType='Created']
* List of authors expressed as tuples that define given name, family name, ORCID. ORCID is a persistent identifier for authors that is optionally available in the XML records. Define the XPath needed to extract such information
* The candidate is free to include further supplementary models supporting the calculations indicated in the next point.

1. Preferably using the Spark framework, calculate the following statistics on the available data

* The number of records per publication year
* The number of records per typology
* The number of Journal Articles produced since 1985, grouped by intervals of 5 years
* Number of records per author
* Number of records per ORCID identified author

1. The project source code must be versioned using git and possibly delivered on Github.
2. Java at least 8, Spark 2.4

[1] [OAI-PMH protocol specification](https://www.openarchives.org/OAI/openarchivesprotocol.html)

|  |  |
| --- | --- |
| Repository base URL | <http://pub.uni-bielefeld.de/oai> |
| Verb for harvesting | ListRecords |
| Metadata prefix | oai\_datacite |

<https://pub.uni-bielefeld.de/oai?verb=ListRecords&metadataPrefix=oai_datacite>