# Problem Statement

The reference loader obtains XML documents from pubmed and gets specific data to be saved to the paper\_citation table in the database.

The most recent reference loader job run has produced the following error report:

DBD::DB2::st execute failed: [IBM][CLI Driver][DB2/LINUXX8664] SQL0433N Value "Brockman</LastName><ForeName>Stephanie A</ForeName><Initials" is too long. SQLSTATE=22001

Unable to execute INSERT INTO PAPER\_CITATION (

object\_paper\_id

, subject\_pubmed\_id

, type\_id

, subject\_title

, subject\_author

, subject\_pub\_year

, CREATED\_BY)

VALUES(

48892

, 22975720

, 65742

, 'The mitochondrial genome of Paraminabea aldersladei (Cnidaria: Anthozoa: Octocorallia) supports intramolecular reco mbination as the primary mechanism of gene rearrangement in octocoral mitochondrial genomes.'

, 'Brockman</LastName><ForeName>Stephanie A</ForeName><Initials>SA</Initials>< AffiliationInfo><Affiliation>Department of Biology, Harvey Mudd College, Claremont, California, USA.</Affiliation></AffiliationInfo></Author><Author ValidYN="Y"><LastName> McFadden'

, '2013'

, 3650

) [IBM][CLI Driver][DB2/LINUXX8664] SQL0433N Value "Brockman</LastName><ForeName>Stephanie A</ForeName><Initials" is too long. SQLSTATE=22001

The Perl script that prepares this data makes a call to fetch the pubmed article using a URL

pmid 34712818

https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed&retmode=xml&id=34712818

reference PMID 22975720

<https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed&retmode=xml&id=22975720>

The code to pull out the data is given here:

Perl code for pattern matching:

($year) = ($text =~ /<Year>(\d+)<\/Year>/);

($lastName) =($text =~ /<LastName>(.+)<\/LastName>/);

($title) =($text =~ /<ArticleTitle>(.+)<\/ArticleTitle>/);

# Task

Determine the root cause of the error. Review the example pubmed XML and describe the nature of the error and recommend remedial action in order to correct the source of the problem.

While reviewing the data that exists in the PAPER\_CITATION table you notice that the “subject\_pub\_year” is not showing the data from the tag you expected to find it from, namely:

<PubDate>

<Year>2021</Year>

</PubDate>

This is happening when the DateRevised happens in a different year.

Describe what you believe is happening when the XML is parsed to obtain the “subject\_pub\_year”.

Prepare a pull request and submit your findings in a word document to github.