How normalized does it have to be? 3nf or is it any form of normalization?

What does the history table contain?

What do you mean by articles in journals?

* Is that just everything between the <Article> start and end tag?
* Are articles within a single journal, but also have the journal tag in between their opening and closing tags?
* Each article belong to one journal?

Questions for Monday August 9th.

Can you use Xpath to return a node?

Yes.

What date really matters?

* There are multiple dates in history, dates under pubmed article, dates under journal and date under article itself?

What do you do about pubdate?

* Lots information missing. Most don’t have day of the month
  + Every medline citation includes two months and no days.
  + Cardinal dates?

We can pick the first month and first day.

Just need to document it.

What is actually supposed to go into the history table? Slightly confused by the teams discussions.

Count the number of revisions per article

Count time between start and finish for the article.

How do authors affiliations work?

* Authors put their affiliations wrong?
  + Does author affiliation matter?
* Do I clean that up and assume that every author with the same last name and initials are the same person?

I’ll just assume that author names are unique.

Is ISSN the unique identifier for an individual journal? If so, why do some with identical ISSN’s have different pub dates?

Need to include volume and Issue for journals.

For a unique Journal, but not a unique journal issue.

Don’t need to store every piece of information for every fact table. Just pick and choose the single fact that you are trying to store and the related information to it.

By journal ISSN. Separate issues of the same journal count as one journal throughout a year.

Don’t need to fill in every field inside a fact table.

Can collapse journal ISSN into fact table.

Max time - Minimum time = days between start and finish

* Or days between received and published.

Author count

Days since article publication

Put the number of articles written by an author in the fact table or the maximum number of articles written by a single author.

Days since last publication.