

Getting Along with Relational Databases

Martin Holmes
University of Victoria
Digital Victorian Periodical Poetry



Outline

Background on relational databases and XML

The DVPP project

Why we ended up with a relational database

Implications, problems and solutions

Relational databases

How do I hate thee? Let me count the ways.

Mysterious and unmemorable keywords (left joins, right joins, inner joins, outer joins, full joins, self joins...)

Lousy string-handling library.

Monolithic single-point-of-failure server host.

No built-in version control.

Ugly structures for one-to-many or many-to-many relationships...

RDBs are fine for simple tabular data

But most Humanities data is not simple and tabular. It's

- deeply nested
- loosely connected
- ambiguous and suggestive
- multivalent

```
...the <date notBefore-custom="1213-05-23"  
notAfter-custom="1214-05-07"  
datingMethod="mol:julianSic"  
calendar="mol:regnal">fifteenth</lb/>  
  yeare of <name ref="mol:JOHN1">King  
    <hi style="font-style:  
italic;">John</hi></name></date>...
```

But the history is revealing

IBM's 1960s DBMS *IMS* represented data as hierarchical trees.

It even handled overlapping hierarchies, allowing "a secondary data structure" which is "still a hierarchy, but a hierarchy in which participant segments have been rearranged, possibly drastically" (C.J.Date, *An Introduction to Database Systems*).

But like so many good ideas from the 1960s...

...it turned into something cruder and less imaginative in the 1970s, with the rise of SQL.

But now we have XML, and life is good, right?

Are you sure you need a database?
Is your data really straightforward?
Wouldn't you be better off with XML?

I'm doing some research and I need a database.

It's really simple.
I just have a single spreadsheet.

Some weeks later...



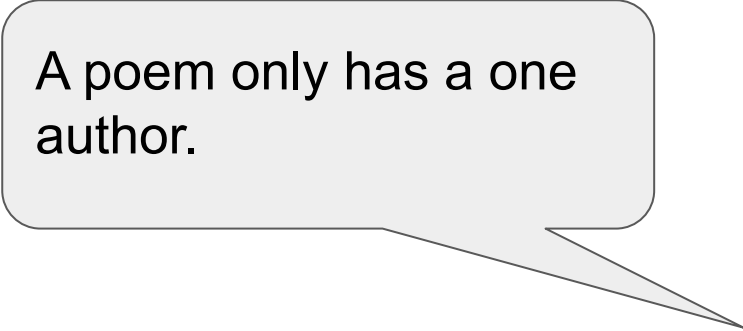
My spreadsheet has
2,517 columns...

Eventually reality bites

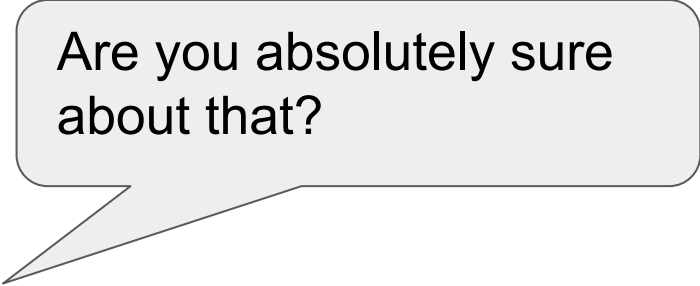
What seems like simple data turns out to be hugely complicated (it's Humanities, doh).

What appear to be simple relationships turn out to be multifarious and nuanced.

What started out as a clean, simple db turns into a monstrous concoction.



A poem only has a one author.



Are you absolutely sure about that?



Absolutely.

Some days later...

This poem has three authors.
One of them is mythical.
Another may be any of three
people. The third is fictional.
What do I enter in the author
field?

The Digital Victorian Periodical Poetry Project

BETA

DIGITAL VICTORIAN PERIODICAL POETRY

"much to do with victorian poetry"



[Transcribed poems](#)

[Poem database](#)

[Personography](#)

[Diagnostics](#)

[Project documentation](#)

[Text analysis](#)

[About us](#)

Welcome!

Digital Victorian Periodical Poetry explores the poetry most read in the long Victorian period: poems published in periodicals, magazines, and newspapers.

This site is in the early stages of development.

The Digital Victorian Periodical Poetry Project

Mines a broad range of 19th-century periodicals for the poetry they published.

Started as a metadata-only project, with poem files linked to page-images.

Data stored in [MySQL DB](#).

Initially simple, but grew in complexity over the years.

Find, enter and edit data

Instructions Tools **Poems** People Organs Hashtags

Poems fields

Search + Show

id

Search

Records to show: Columns to show: Pages: Total found: 1

	id	Title	First line	Authors	Translators	Date	Organ	Series	Actions	*
<input type="checkbox"/>	10001	The Master	The herd of Scribes by what they tell us	Schiller, Johann Christoph Friedrich von	Bulwer-Lytton, Edward George	1843-03	Blackwood's Edinburgh Magazine		<input type="button" value="▼"/> <input type="button" value="✎"/> <input type="button" value="✕"/>	

id **10001**

Title **The Master**

First line **The herd of Scribes by what they tell us**

Authors **Schiller, Johann Christoph Friedrich von**

Translators **Bulwer-Lytton, Edward George**

Date **1843-03**

Organ **Blackwood's Edinburgh Magazine**

Series

Vol. **53**

Num. **329**

Pages **310**

Quoted in article

Allonymous?

Allonym

Unsigned?

Then we got funding

Thank you SSHRC!



Social Sciences and Humanities
Research Council of Canada

Conseil de recherches en
sciences humaines du Canada

Canada 

We're transcribing/encoding all the poems from the decade-years (1920, 1930, 1940...).

This will be around 2,000 poems.

We're encoding in TEI (of course).

We're using a version-control system (of course).

So now we have a problem.



Metadata db:

Canonical source for most metadata

Continuously changed and updated



TEI files:

Canonical source for transcription and some metadata

Continuously changed and updated

Added complication

We have a personography for which some fields of some records are maintained in the db, and some fields and some records are maintained in the TEI.

Luckily this is all temporary

Eventually everything will be maintained in TEI.

We will bury the database.

We will stamp on its grave and rejoice.

But for now, we have to live with it.

The task

Merge the metadata from the db into the TEI file collection.

Without data loss.

Repeatedly.

The ~~only good~~ best thing about SQL Dbs is...

... you can dump them to XML!

Dump DB
to XML

Dump DB
to XML



Rebuild
TEI historical
personography

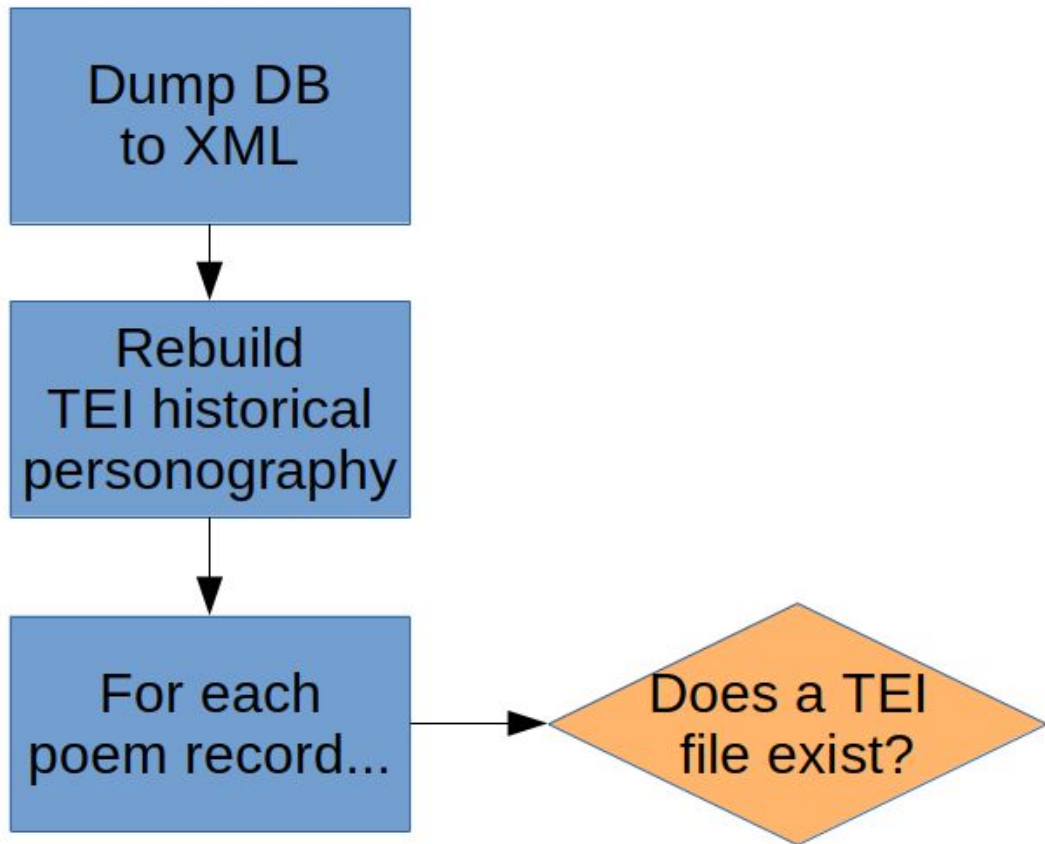
Dump DB
to XML



Rebuild
TEI historical
personography



For each
poem record...

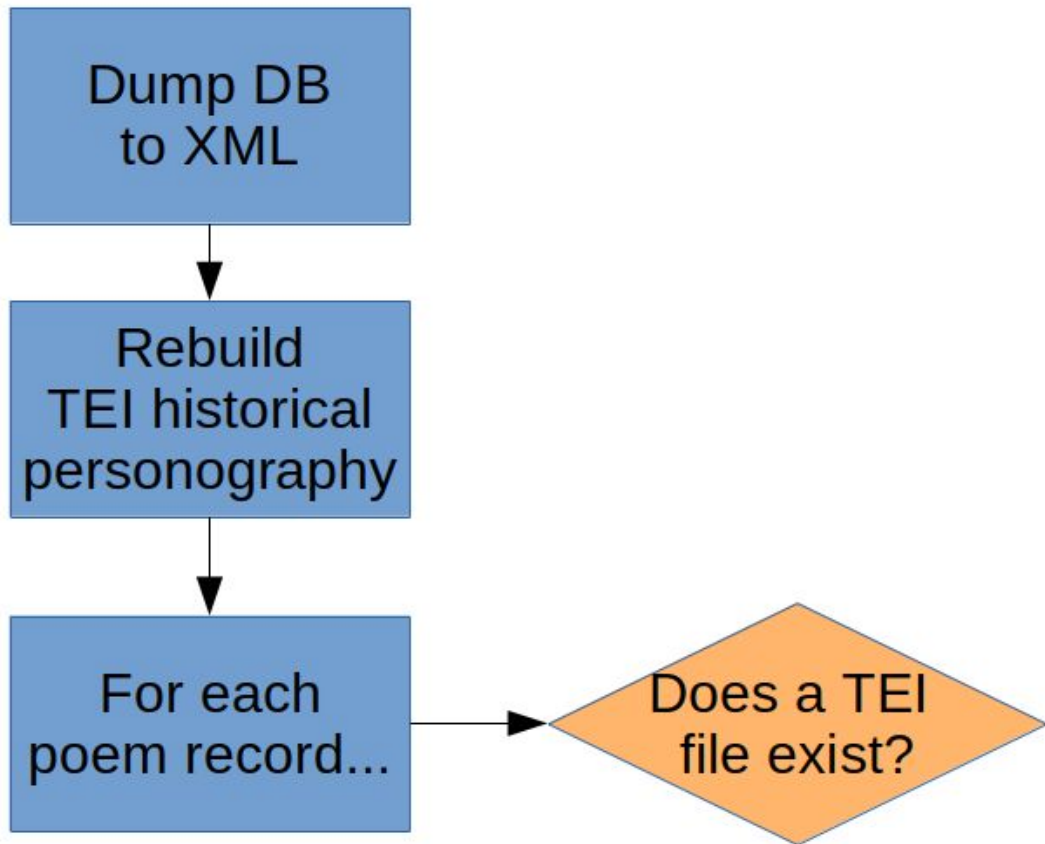


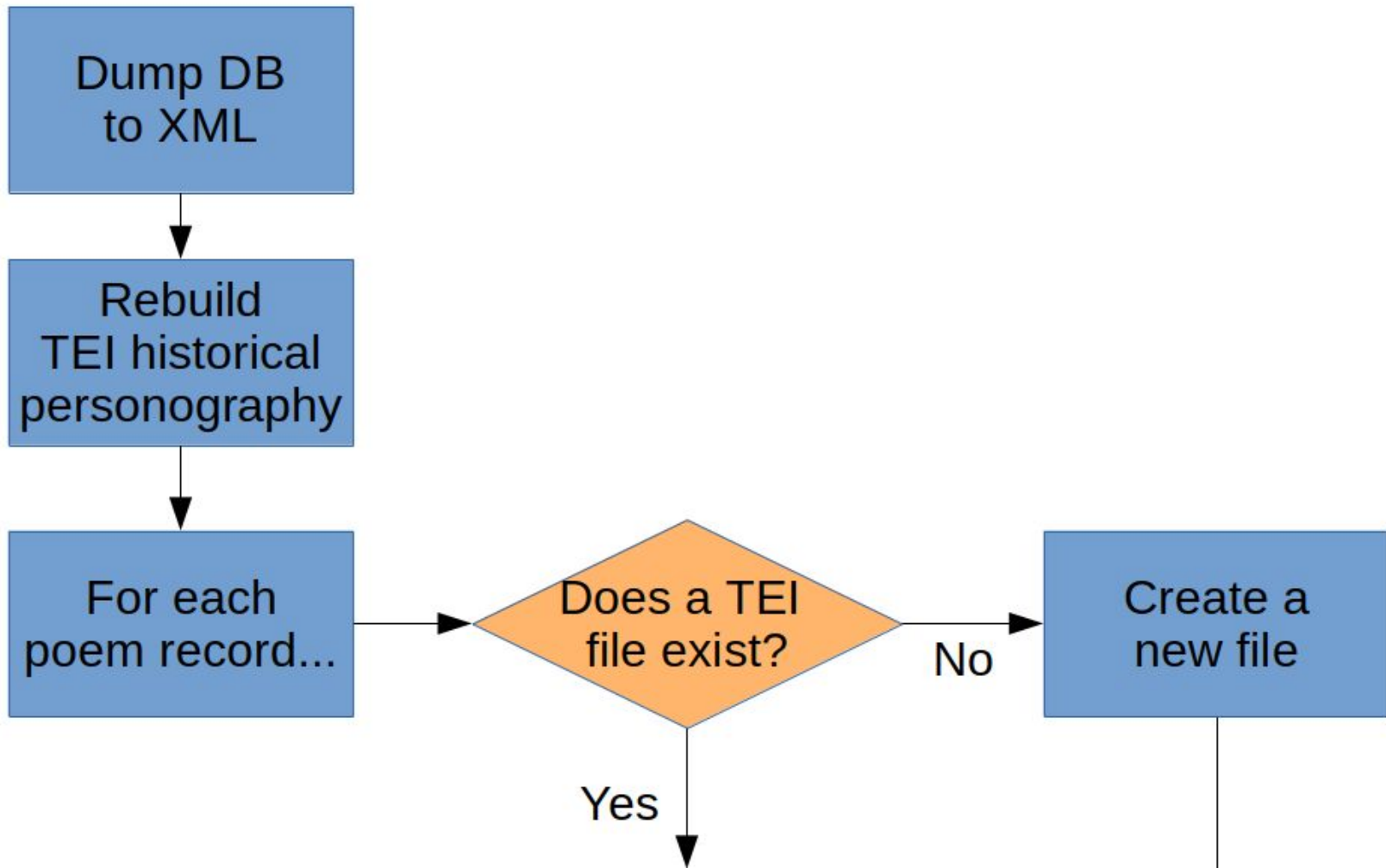
TEI Poem ids/filenames

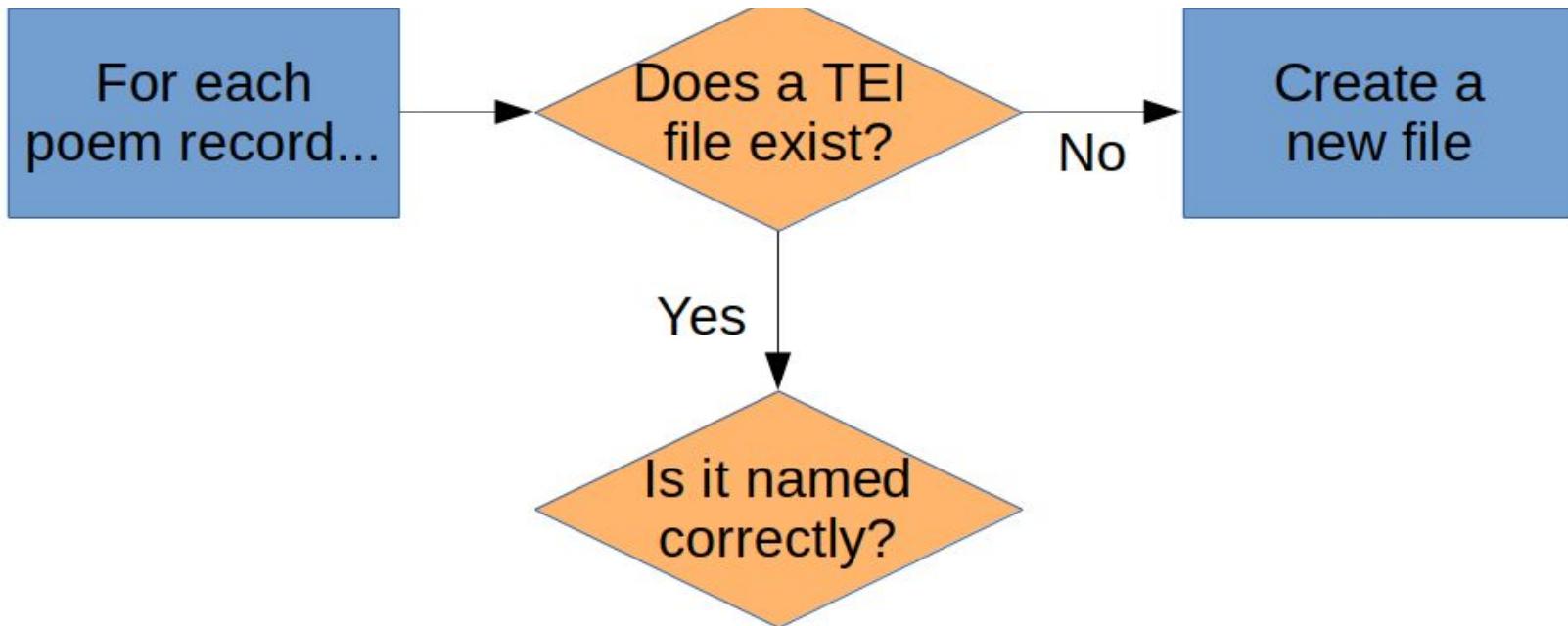
Derived from db record id and poem title:

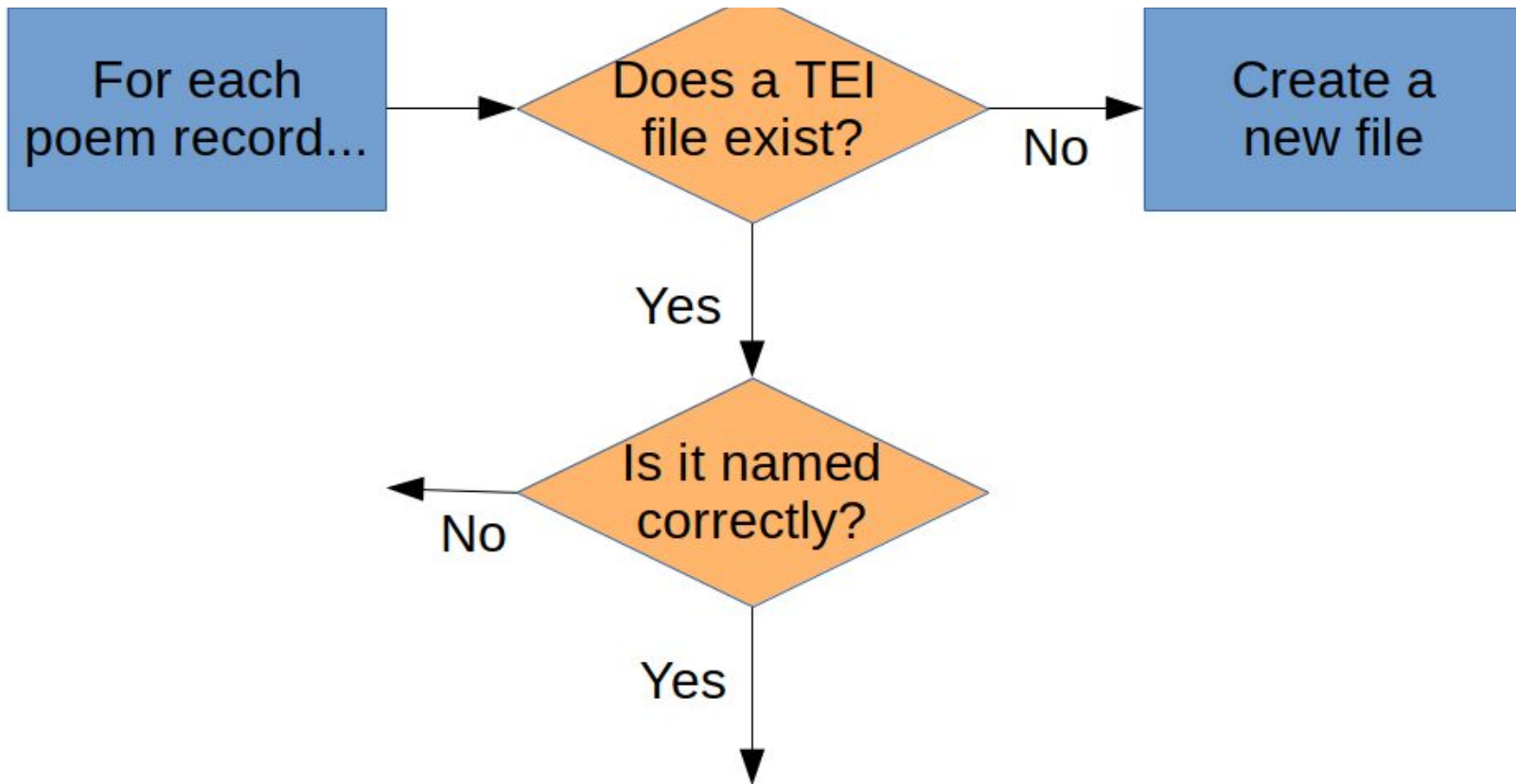
pom_7824_reflections_on_a_brumel_scene

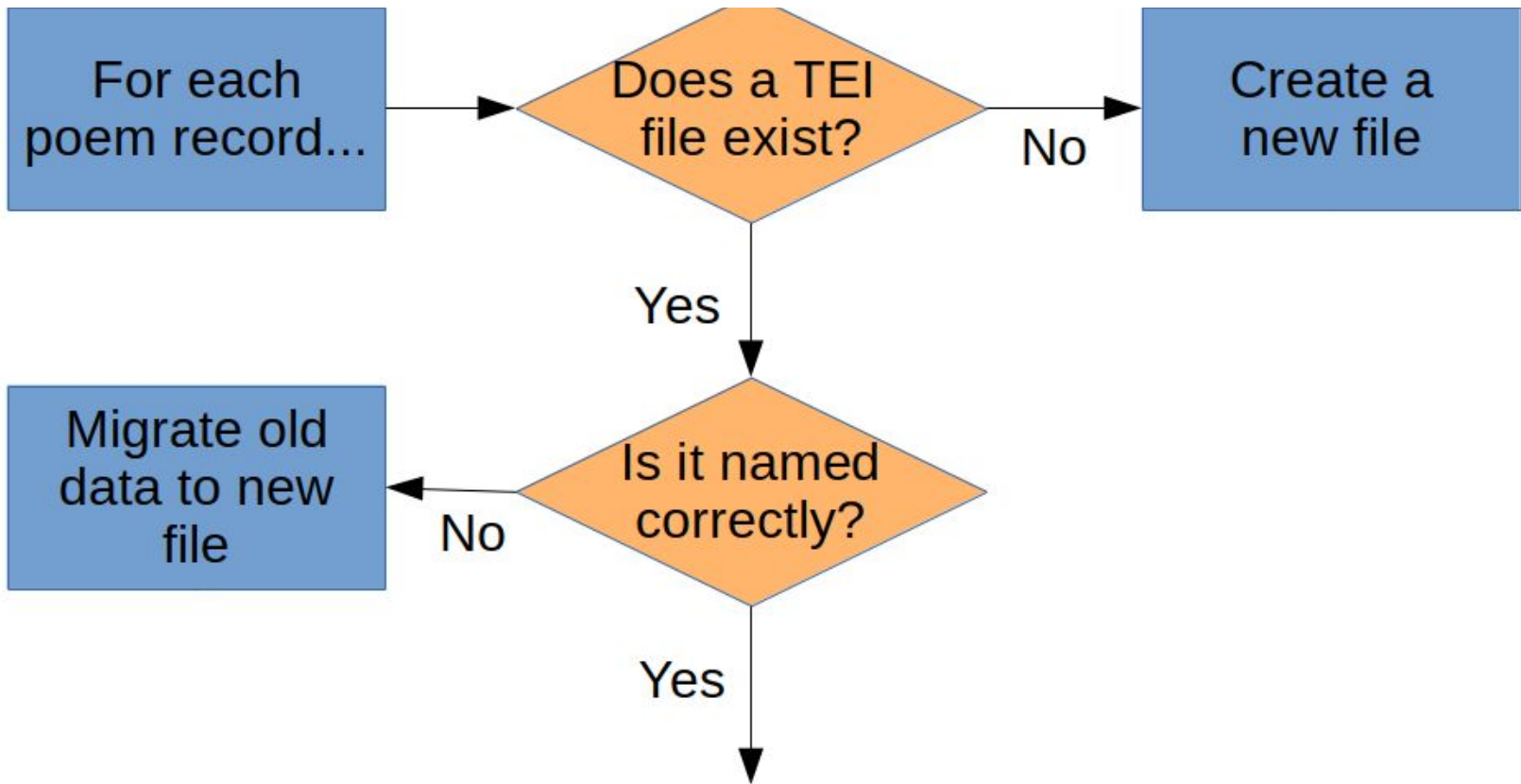
Filenames always match root ids + .xml.

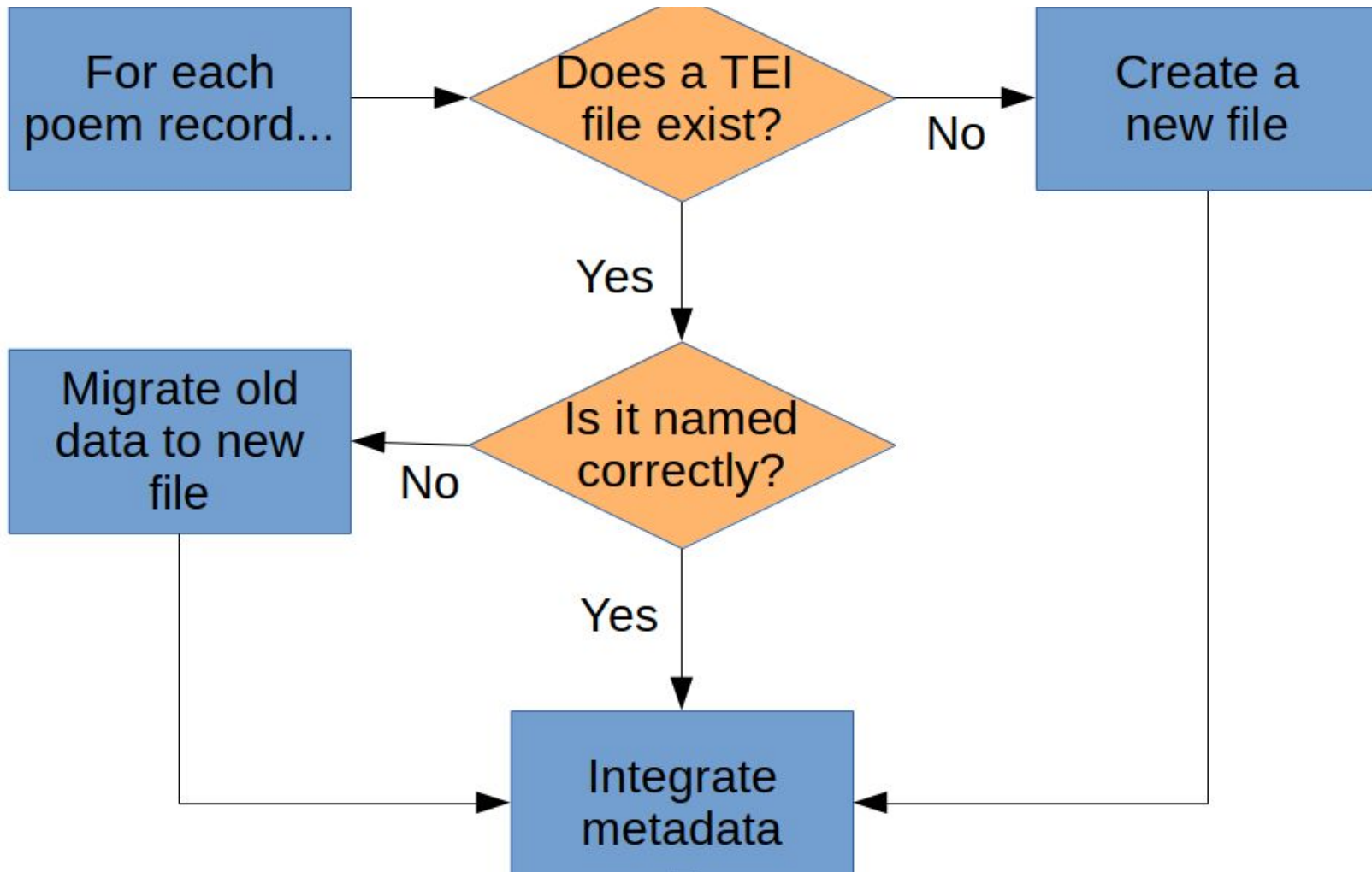




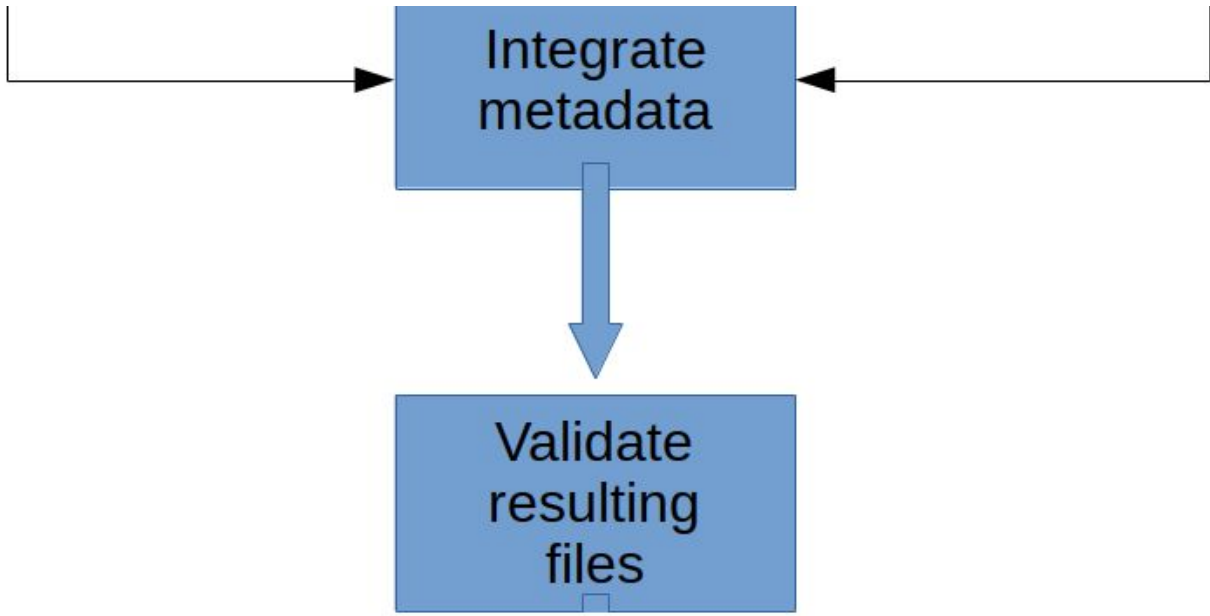


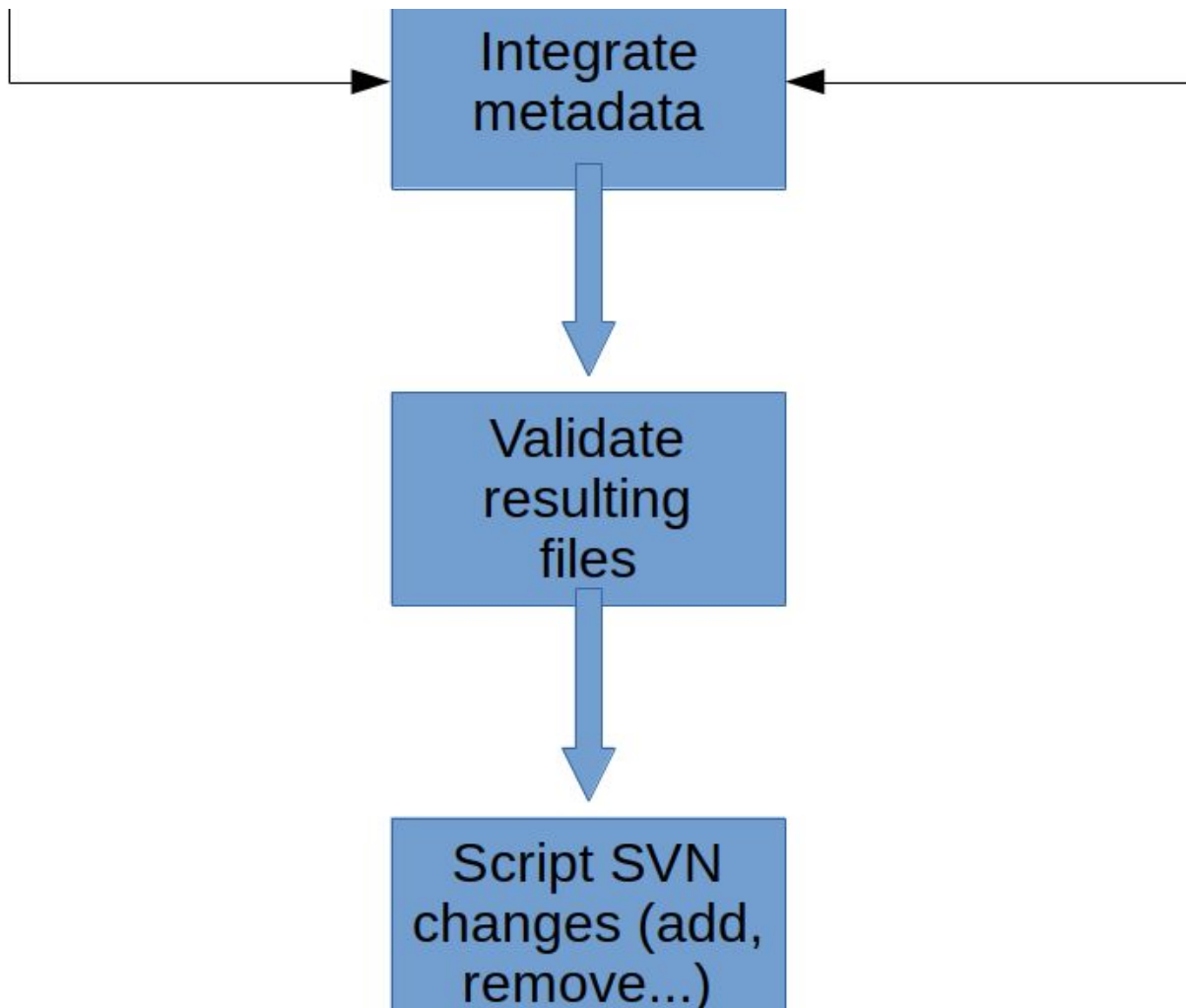












SVN Changes

- If a file is **new** and **not tracked by svn**, it must be **added**.
- If a file has **not changed** during this operation, then it is **obsolete** and must be **removed** from svn.
- If a file has **changed** during this operation, it can simply be **committed**.

Result

The two teams (poem harvesters and TEI encoders) can work in parallel.

Periodic application of the integration process keeps the two datasets in sync.

TEI encoding is not in danger of being overwritten.

A few more scenarios where TEI wins

Two pseudonyms may or may not represent the same author.

Some poems claim to be translations but are probably not; their “translators” are most likely their authors.

Some pseudonyms represent changing teams of anonymous authors, only some of whom we can identify.

Interim strategy: hashtags in the db

Without complexifying its structure, the db can't handle weird cases like this. But we can flag them for later.

For this, we use *hashtags* in Notes fields.

Hashtags fields

Search + Show

Search Spreadsheet

Records to show: 20 Columns to show: 4 Pages: 1 Total found: 14

<input type="checkbox"/>	id	Hashtag	Gloss	Description	Actions *
<input type="checkbox"/>	14	#abbrevNLS	fix NLS entry for abbreviated note	shortcut for NLS notes, which need amending for previous entries in DB poem note field, such as an address mentioned	<input type="checkbox"/> ▼
<input type="checkbox"/>	1	#badCitation	Bad citation	One or more citations needs to be checked and corrected.	<input type="checkbox"/> ▼
<input type="checkbox"/>	13	#fauxTranslation	Faux Translation	A poem that falsely claims to be a translation.	<input type="checkbox"/> ▼
<input type="checkbox"/>	2	#fixDates	Fix dates	One or more dates in the record needs to be checked and corrected; or the dates are BCE or approximate	<input type="checkbox"/> ▼
<input type="checkbox"/>	3	#illustratorIllegible	Illustrator illegible	The identity of the illustrator is not clear because of illegible text.	<input type="checkbox"/> ▼
<input type="checkbox"/>	4	#makeRelatedPoems	Make related poems	This poem is related to other poems in the index, and the relation fields have not yet been created and/or the relation needs to	<input type="checkbox"/> ▼
<input type="checkbox"/>	5	#missingCitation	Missing citation	One or more citations that should appear in the record are missing.	<input type="checkbox"/> ▼

Illustrations

Links

Notes **Included in a series of poems with the general title "From the Sicilian of Vicortai," unsigned by the translator (742-6) (KF). Translator attribution: James Logie Robertson, *Orellana, and other Poems*, William Blackwood and Sons, 1881, pp. 125-43. These translations are in fact original poems by Robertson; see David Herschel Edwards, editor, *One Hundred Modern Scottish Poets*, third series, D. H. Edwards, 1881, p. 354. (AC).**

Hashtags **#fauxTranslation**

Use of hashtags cannot be constrained in the db.

However, we can validate it during the build process by creating and applying diagnostic rules based on the Hashtags table.

We can also generate a hashtag taxonomy in the TEI and convert usages into <catRef> elements.

Links

The (nascent) project website: <https://dvpp.uvic.ca>.

The XSLT for merging metadata into poems:

https://hcmc.uvic.ca/svn/dvpp/xsl/sql_to_tei_master.xsl

The Ant task that runs the process:

<https://hcmc.uvic.ca/svn/dvpp/buildTEI.xml>

The project documentation on this process:

<https://dvpp.uvic.ca/dvpp.html#refreshDatabases>