EDIT - bug #9386

orthographic variants should not be symmetrical

01/13/2021 06:12 PM - Andreas Müller

Status:	New	**Start date:**	
Priority:	Highest	**Due date:**	
Assignee:	Andreas Müller	**% Done:**	0%
Category:	cdm	**Estimated time:**	0:00 hour
Target version:	CDM UML 5.43		
Severity:	major	**Found in Version:**	

Description

According to the nomenclatural code there is always only 1 valid orthographic variant so the name relationship should always direct from the invalid variant to the valid variant and therefore the relationship should not be symmetrical.

https://www.iapt-taxon.org/nomen/pages/main/art_61.html

====

List of usecases (to be extended):

1. Search by invalid variants to find the valid variants (much easier or only possible if relation is asymmetric)
2. Show original spelling for a name "[as xxx]" (a specific usecase of the asymmetric relation, but handled in the existing specific relation "is original spelling for ", so not relevant here)
3. If an invalid variant is also included in the synonymy show what the valid variant of the "name" is (much easier or only possible if relation is asymmetric)
4. Show all variants of the name (easier if relation is asymmetric but asymmetry is not required; in general this usecase has never been a requirement of any user in the past)
5. Automatic replacement of invalid names being orthographic variants with the validly published form of that name as required by the code (§61.4) (much easier or only possible if relation is asymmetric)
6. Data validation/consistency check (more profound and easier if relation is asymmetric)
7. ...

Related issues:

Related to EDIT - task #7960: Revision of NameRelationshipTypes	**New**
Related to EDIT - discussion #6678: [DISCUSS] How to correctly show name rela...	**In Progress**
Related to EDIT - bug #10311: Is misspelling for and is emendation for should...	**Closed**
Blocks EDIT - feature request #4640: Add orthographic variants to name & taxa...	**New**

History

#1 - 01/13/2021 06:29 PM - Wolf-Henning Kusber

It is correct that there is only one correct spelling. In PhycoBank, we need to register the published name "as is" = the valid name but incorrect spelling, whereas the correction is done with help of the Code, normally without a reference. Both spellings should be searchable in the System.

#2 - 01/14/2021 10:12 AM - Andreas Müller

WGB:

richtig, Art. 61.1 sagt, dass es nur eine gültige („valid“, nicht „correct“) Variante gibt, aber er spricht auch ausdrücklich von den Varianten, von denen nur eine gültig sein kann. Also ist die gültige auch eine Variante der ungültigen. Die ungültigen Varianten können als nom. inval. gekennzeichnet werden. Also finde ich die Beziehung eigentlich in Ordnung, wie sie ist.
Wenn das wie vorgeschlagen geändert wird, dann sollte auch die Bezeichnung der Beziehung entsprechend geändert werden („is invalid orthographic variant of“) – aber das ist eigentlich eine unnötige Duplizierung der Statusangabe, finde ich.

#3 - 01/14/2021 10:16 AM - Andreas Müller

Wolf-Henning Kusber wrote:

> It is correct that there is only one correct spelling. In PhycoBank, we need to register the published name "as is" = the valid name but incorrect spelling, whereas the correction is done with help of the Code, normally without a reference. Both spellings should be searchable in the System.

Please be aware that this ticket is about relation "is orthographic variant for". Of course original spellings are also orthographic variants but special ones and usually we do not store them with the above relation but with "is original spelling for" or with the latest release simply in the

NameUsedInSource/Original Spelling field in the nomenclatural source.
So we discuss here only orthographic variants which are NOT used in the original publication but in other publications.

Searching is another issue (which is unfortunately not yet implemented correctly).

#4 - 01/15/2021 09:50 AM - Andreas Müller

ERS:

ja, es kann nur eine korrekte orthographische Variante geben, und daher ist die Richtung nicht symmetrisch: siehe Artikel 61 Shenzhen Code.
https://www.iapt-taxon.org/nomen/pages/main/art_61.html

Mit anderen Worten, es handelt sich bei einer Beziehung zwischen einem Namen und seiner orthographischen Variante nicht um eine Beziehung zwischen zwei gleichermaßén korrekten Namen (wie etwa eine Basionym – comb.nov.-Beziehung oder ähnliches).

#5 - 01/15/2021 10:36 AM - Andreas Müller

> WGB:
>
> richtig, Art. 61.1 sagt, dass es nur eine gültige („valid“, nicht „correct“) Variante gibt, aber er spricht auch ausdrücklich von den Varianten, von denen nur eine gültig sein kann. Also ist die gültige auch eine Variante der ungültigen. Die ungültigen Varianten können als nom. inval. gekennzeichnet werden. Also finde ich die Beziehung eigentlich in Ordnung, wie sie ist.
> Wenn das wie vorgeschlagen geändert wird, dann sollte auch die Bezeichnung der Beziehung entsprechend geändert werden („is invalid orthographic variant of“) – aber das ist eigentlich eine unnötige Duplizierung der Statusangabe, finde ich.

So we have 2 different handlings of this relationship now which emphasizes the unclear semantics of the relationship which is caused by it's label.
So I guess it is good to make the semantics clearer.

First we need to decide if we want to have only 1 relationship being either symmetric or asymmetric or if we want to have 2 relationships 1 being symmitric the other one assymmetric.

What are the pros and cons for theses relationships:

pro asymmetric: more exact in terms of the code Art. 61; easier to compute data for search or to validate as the only correct data schema is a directed star schema here while the symmetric version allows complex graphs which are difficult to compute.

pro symmetric: legacy data sometimes is not clear about the direction so automated imports might be difficult as taxonomic knowledge is needed to decide

Any other pros and cons?

#6 - 01/15/2021 10:46 AM - Wolf-Henning Kusber

What is about data exports? Isn't it maybe more problematic if the relation symetrical in EDIT?

#7 - 01/15/2021 11:00 AM - Andreas Kohlbecker

Hallo,

auch wenn diese Beziehung zwischen zwei Namen im Modell als ungerichtet definiert ist, besteht dennoch eigentlich immer eine implizite Richtung, die sich letztlich aus dem Publikationsdatum des einen und des anderen Namen ergibt.
Der Name neueren Datums ist (immer?) der korrigierende Name, der auf den invaliden Namen verweist. Oder gibt es Fälle in denen eine Autor in der irrigen Annahme, ein Name sei invalide, diesen ungültig korrigiert - es kommt ja so manches vor - und das Publikationsdatum ist daher keine verlässliche Angabe? Wie oft kann so ein Fall vorkommen? Würde diese Inferenz nicht auch das von Andreas M. genannte Problem bei Importen lösen? In nicht auflösbaren Fällen bliebe ja noch immer die Möglichkeit der "Annotation", um so auszudrücken, dass die Richtung der Beziehung nicht wirklich bekannt ist.

Ich würde dafür plädieren die Benennung der Relation beizubehalten, sie aber im Modell als explizit gerichtet zu definieren.

Viele Grüße,
Andreas

#8 - 01/15/2021 11:23 AM - Wolf-Henning Kusber

Das Datum sollte die meisten Fälle korrekt einsammeln. Wie Andreas sagt, sollte eine Richtung erkennbar sein. Ausnahmen gibt es immer, wie es auch Zweifelsfälle gibt. Deshalb sollte "inval." in der Beziehung nicht genannt werden (wie Walter eingangs überlegt hatte) und ein invalider oder illegitimer Name entsprechend nach Evaluiertung einen entsprechenden Status bekommen.

#9 - 01/15/2021 11:54 AM - Andreas Müller

Andreas Kohlbecker wrote:

> auch wenn diese Beziehung zwischen zwei Namen im Modell als ungerichtet definiert ist, besteht dennoch eigentlich immer eine implizite Richtung, die sich letztlich aus dem Publikationsdatum des einen und des anderen Namen ergibt.
> Der Name neueren Datums ist (immer?) der korrigierende Name, der auf den invaliden Namen verweist. Oder gibt es Fälle in denen eine Autor in der irrigen Annahme, ein Name sei invalide, diesen ungültig korrigiert - es kommt ja so manches vor - und das Publikationsdatum ist daher keine verlässliche Angabe? Wie oft kann so ein Fall vorkommen? Würde diese Inferenz nicht auch das von Andreas M. genannte Problem bei Importen lösen? In nicht auflösbaren Fällen bliebe ja noch immer die Möglichkeit der "Annotation", um so auszudrücken, dass die Richtung der Beziehung nicht wirklich bekannt ist.

Es geht hier wie gesagt NICHT um die Korrektur von original spellings sondern um SONSTIGE orthographic variants. Daher verstehe ich nicht, warum das Datum der validen Variante immer später liegen muss. Werden orthographische Varianten nicht eher in sonstigen Publikationen verwendet wo jemand den Namen verwendet aber eben in einer invaliden Variante. Die kann doch dann genauso später liegen als die valide publizierte Variante. Noch dazu verlangt das, dass das Datum für beide Namen immer bekannt und in der Daten vorhanden ist, was ebenfalls längst nicht immer der Fall ist.

#10 - 01/15/2021 12:07 PM - Andreas Müller

Andreas Kohlbecker wrote:

> Ich würde dafür plädieren die Benennung der Relation beizubehalten, sie aber im Modell als explizit gerichtet zu definieren.

Aber dann sagt das Label etwas anderes aus als das Modell, wie die Misinterpretation von mir und Walter zeigen, die das Label so interpretiert haben, dass es eine symmetrische Beziehug sein soll. Das finde ich nicht sinnvoll. Das Label sollte, wenn möglich die Semantik abbilden.

#11 - 01/15/2021 12:10 PM - Andreas Müller

- File picture616-1.png added

Andreas Kohlbecker wrote:

> auch wenn diese Beziehung zwischen zwei Namen im Modell als ungerichtet definiert ist, besteht dennoch eigentlich immer eine implizite Richtung, die sich letztlich aus dem Publikationsdatum des einen und des anderen Namen ergibt.
> Der Name neueren Datums ist (immer?) der korrigierende Name, der auf den invaliden Namen verweist. Oder gibt es Fälle in denen eine Autor in der irrigen Annahme, ein Name sei invalide, diesen ungültig korrigiert - es kommt ja so manches vor - und das Publikationsdatum ist daher keine verlässliche Angabe? Wie oft kann so ein Fall vorkommen? Würde diese Inferenz nicht auch das von Andreas M. genannte Problem bei Importen lösen? In nicht auflösbaren Fällen bliebe ja noch immer die Möglichkeit der "Annotation", um so auszudrücken, dass die Richtung der Beziehung nicht wirklich bekannt ist.
>
> Ich würde dafür plädieren die Benennung der Relation beizubehalten, sie aber im Modell als explizit gerichtet zu definieren.

WGB:

ich denke, es gibt eine Menge Fälle, in denen eine (nach Code invalide) orthographic variant später gebildet wird, z.B. durch eine Verschlimmbesserung der Endung auf Grund eines falsch angenommenen Geschlechts des Gattungsnamens. Der Name hat dann die gleichen Publikationsdaten wie der valide Name. Der Fall ist analog zu einer Sache, die wir beim Sammlungs-Modell 1999 und später, glaube ich, auch in ABCD als Assemblage definiert hatten – es handelt sich eigentlich nicht um eine Beziehung zwischen 2 Objekten, sondern eine Aggregation mehrerer Objekte:

> While the association defines an asymmetric relationship between units (mostly pairs of units), the assemblage serves to handle symmetric relationships, mostly among several units, according to a single criterion ("Assemblage name", "Assemblage description"). A typical example is a herd of animals, either in the wild or in a zoological garden, or any kind of symmetric symbiontic, mutualistic or commensalistic relationship . Other applications of assemblages are batches of eggs taken

#12 - 01/15/2021 12:17 PM - Andreas Müller

Wolf-Henning Kusber wrote:

> What is about data exports? Isn't it maybe more problematic if the relation symetrical in EDIT?

This depends on the target format but in general it is true that an assymetric relation can be mapped more exact to export formats

#13 - 01/15/2021 12:29 PM - Andreas Müller

Ich möchte nochmal explizit auf die Probleme hinweisen, die sich datenverabeitungsseitig aus symmetrischen Beziehungen ergeben. Da diese in aller Regel nicht vollständig eingegeben werden (jede Variante wird mit jeder verknüpft) sondern nur so dass die Namen irgendwie verknüpft sind, kann es deutlich aufwendiger werden auch nach Namensvarianten zu suchen um dann den zugehörigen validen Namen im Portal anzuzeigen. Im Extremfall von 10 Varianten müsste die linear miteinander verknüpft sind, müsste ich mich durch 9 Namensrelationen hangeln bis ich zum validen Namen komme, was unschön ist. Eine Erhöhung der Komplexität kann nur vermieden werden, wenn eine Vorberechnung des vollständigen Graphen regelmäßig erfolgt, was Implementierungsaufwand bedeutet.
Bei einer asymmetrischen Beziehung ist (bei korrekten Daten) vorgegeben, dass alles in Sternform miteinander verknüpft sein muss, was die

Auswertung deutlich leichter macht.

Dieses Argument ist für mich das Hauptargument, warum ich dafür bin die gerichtete Relation zu verwenden, auch wenn es dadurch Probleme mit unvollständigen Daten beim Import geben kann. In anderen Fällen bin ich ja ein großer Freund dessen, dass solche unvollständigen oder auch dirty Data bei Importen berücksichtigt werden können müssen, in diesem Fall überwiegt für mich aber das erste Argument, da ich auch glaube, dass das Importproblem nicht so häufig auftritt und dann eben mit Logfiles, Annotations etc. gelöst werden muss.
Anders wäre es allerdings, wenn es taxonomisch wirklich schwer zu entscheiden wäre, welches die valide Variante ist.

#14 - 01/15/2021 01:21 PM - Andreas Kohlbecker

[Eckhard wrote]

Hallo,

nach neuerlichem Lesen des At. 61 gebe ich dir recht, Walter: die gültige orthographische Variante ist eben auch eine Variante der ungültigen. Das hatte ich vorher intuitiv anders gesehen, ich dachte, ein gültiger Name hätte eine orthographische Variante. So gesehen könnte ich also auch für einen gültigen Namen die Beziehung erstellen "is orthographic variant of", das war mir bisher nicht klar gewesen.

Man muss mindestens zwei Fälle unterscheiden:

1.) die ursprüngliche, ältere Schreibweise ist die richtige, z.B. Limonium longibracteatum. Dazu gibt es eine spätere (invalide) orthographische Variante, also etwa Limonium longebracteatum. (das ist aber m.E. *kein* nom. inval., sondern überhaupt kein anderer Name, nur eine falsche Schreibweise des ursprünglichen Namens ohne neuen Autor, ohne neues Jahr). Diese Variante war nur als vermeintliche Verbesserung der Originalschreibweise gedacht, was in dem Fall nicht richtig war, aber nicht als neuer Name.

Sehr häufig wird ein Name in einer späteren Veröffentlichung schlicht und einfach falsch geschrieben. Manchmal setzt sich diese falsche Schreibweise in vielen Publikationen fort.

Die Beziehung ist daher vielleicht insofern nicht symmetrisch, denke ich, als der ältere Name gültig ist, die spätere orthographische Variate aber überhaupt kein anderer Name, sondern eine ungültige Variante *desselben* Namens.

Vgl. Art. 61.4: The orthographical variants of a name are to be corrected to the validly published form of that name. Whenever such a variant appears in a publication, it is to be treated as if it appeared in its corrected form.

Hervorhebung von mir: es sind "Variants of a name", not two different names!

2.) die ursprüngliche, ältere Schreibweise ist die falsche (ein correctable error), und die korrigierte Form die richtige. Die korrigierte Form des Namens wird mit derselben Autorschaft und derselben Jahreszahl wie die urspünglich zitiert, da es ja ein und derselbe, gültig publizierte Name ist, nur eben in einer korrigierten Variante. Wer und wann da korrigiert hat, wird üblicherweise nicht festgehalten. Dieser Fall ist aber durch die Möglichkeit der Eingabe des Original spelling (komplett?) abgedeckt.
Aber anscheinend wird bei der Eingabe eines original spelling im TaxEditorkeine Namensrelation hergestellt ("is misspelling for..." gibt es ja als Relation).

Man könnte die ursprüngliche Form aber auch als orthographic variant festhalten. Aber das ist anscheinend nicht beabsichtigt? Im Berlin Model hatten wir das glaube ich so gemacht...

Wichtig dabei ist ja aus meiner Sicht vor allem, dass der Nutzer bei einer Suche alle möglichen orthographischen Varianten eingeben kann und dann immer zum richtigen Namen geleitet wird.

Ich hoffe, damit nicht zu weiterer Verwirrung beizutragen...

Viele Grüße

Viele Grüße, Eckhard

#15 - 01/15/2021 01:31 PM - Andreas Kohlbecker

Hallo,

ich denke mittlerweile, dass wir zwei Beziehungen brauche wie schon von Andreas Müller vorgeschlagen:

> First we need to decide if we want to have only 1 relationship being either symmetric or asymmetric or if we want to have 2 relationships 1 being symmitric the other one assymmetric.

Wobei die gerichtete Relation zwischen *echten* Namen bestehen kann und immer dann verwendet werden soll wenn ein invalider Name im Sinne des Codes korrigiert wird. Als Label dieser Relation: "Invalid name correction"

Die ungerichtete Relation wäre dann für solche Fälle verwendbar, wie sie Eckhard beschrieben hat. Als falsche Schreibweisen und Varianten die aber keinen echten neuen Namen erzeugt haben. Ach im Falle von unklaren Verhältnissen bei Importen wäre diese Relation dann verwendbar. Für diese Relation könnte das label "Orthographic variant" einfach beibehalten werden.

#16 - 01/15/2021 11:06 PM - Andreas Müller

- Description updated

#17 - 02/01/2021 09:27 PM - Andreas Müller

- Related to task #7960: Revision of NameRelationshipTypes added

#18 - 02/02/2021 10:20 AM - Andreas Müller

- Related to feature request #4640: Add orthographic variants to name & taxa search added

#19 - 02/02/2021 10:23 AM - Andreas Müller

So I think we can summerize that we agree that according to the code orthographic variants are a cluster of "names" where between each of these "names" we could have a relation "is orth. var. for" in both directions. Therefore the existing symmetrical relation is not wrong.
At the same time according to the code only 1 of these names is valid and therefore also to define an asymmetric relation between the invalid names and the 1 valid name is not incorrect or against the code.

So if both is possible the more important question is what is needed and why the relation should be stored in a database at all.
Also the question how to label each of these relations should be discussed separately after the decision was made which relationship we want to have.

To decide what is needed we should look at the usecases that use the relation.
The following (extendable) list tries to list all relevant usecases:

1. Search by invalid variants to find the valid variants (much easier or only possible if relation is asymmetric)
2. Show original spelling for a name "[as xxx]" (a specific usecase of the asymmetric relation, but handled in the existing specific relation "is original spelling for ", so not relevant here)
3. If an invalid variant is also included in the synonymy show what the valid variant of the "name" is (much easier or only possible if relation is asymmetric)
4. Show all variants of the name (easier if relation is asymmetric but asymmetry is not required; in general this usecase has never been a requirement of any user in the past)
5. Automatic replacement of invalid names being orthographic variants with the validly published form of that name as required by the code (§61.4) (much easier or only possible if relation is asymmetric)
6. Data validation/consistency check (more profound and easier if relation is asymmetric)
7. Any other usecase?

At the same time we have to have in mind the ease of data editing. In general it is of course easier to create just any symmetric orth. var. relationship instead of creating the asymmetric relationship only because

1. one does not have to think about the correct direction,
2. automated data imports do not have to think about which name is the valid one
3. the relation may be created between any of the variants and should not necessarily be created between an invalid and the valid variant only

However, as orth. var. relations are usually created during editing of classifications and synonymies it is almost always clear from the context which name is the valid one.

Also there is the argument that one can retrieve the correct direction of the relationship by checking the nomenclatural status of the 2 names. Unfortunately this is not true, because this is not only more difficult to implement and bad for performance but in almost all cases this information is not stored in databases. I checked existing databases for availability (0/41 in caryo_spp, 101/1994 in cichorieae, 0/3 in cyprus, 155/3966 in E+M, 0/4 in flora-of-cuba).
The reason, why this information is usually not available is, that invalid names being orth. variants do usually not show up in synonymies explicitly. So there is not need in adding the status to the name. However, by using an asymmetric relationship, the nom. inval. information could be automatically added to the name without an additional action required.

From the above I strongly recommend to introduce an asymmetric relationship as it is the only one that really covers the required use-cases. Keeping only the existing symmetric relationships will either make writing SQL-like queries for the given usecases extremely cumbersome or even impossible if the nom. status can not be retrieved somehow.
This is also one of the reasons why e.g. searching for orthographic variants has not yet been implemented though it was a requirement for long time already: #4640.

So, those who plead for a symmetric relationship, can you please give the usecases you have in mind and tell why they are better or similar covered by a symmetric relationship?

#20 - 02/02/2021 10:30 AM - Andreas Müller

- Related to deleted (feature request #4640: Add orthographic variants to name & taxa search)

#21 - 02/02/2021 10:31 AM - Andreas Müller

- Related to deleted (feature request #4640: Add orthographic variants to name & taxa search)

#22 - 02/02/2021 10:31 AM - Andreas Müller

- Blocks feature request #4640: Add orthographic variants to name & taxa search added

#23 - 02/02/2021 10:43 AM - Andreas Müller

- Description updated

#24 - 02/03/2021 03:39 PM - Andreas Müller

WGB:

ich bin nur nicht damit einverstanden, dass man das existierende Label einfach asymmetrisch macht. Das klang so – also alles klar.

#25 - 02/03/2021 03:44 PM - Andreas Müller

AM:

nein, nur die bisherige Beziehung umdeuten ist natürlich nicht geplant.

Eher sollte es so aussehen:

• Es wird eine neue, asymmetrische Relation erzeugt
• Die alte Relation wird auf deprecated gesetzt und steht zum editieren neuer Daten nicht mehr zur Verfügung
• Die Daten werden nach Rücksprache mit den Projektverantwortlichen migriert, weitgehend kann das automatisiert erfolgen (die Namen, die in Synonymien stehen, sollten die validen sein, wenn genau einer von beiden Namen in einer Synonymie steht, zudem können existierende Status verwendet werden). Die restlichen Daten werden manuell bearbeitet
• Löschen der alten Relation (außer ein Projekt will sie unbedingt parallel noch behalten)

#26 - 02/06/2021 06:47 PM - Andreas Müller

- Related to discussion #6678: [DISCUSS] How to correctly show name relationship "orth. var." in dataportal added

#27 - 04/27/2023 04:12 PM - Andreas Müller

- Related to bug #10311: ls misspelling for and is emendation for should be asymmetric added

Files

picture616-1.png	80.5 KB	01/15/2021	Andreas Müller