
Learning How to Correct a Knowledge Base from the Edit History

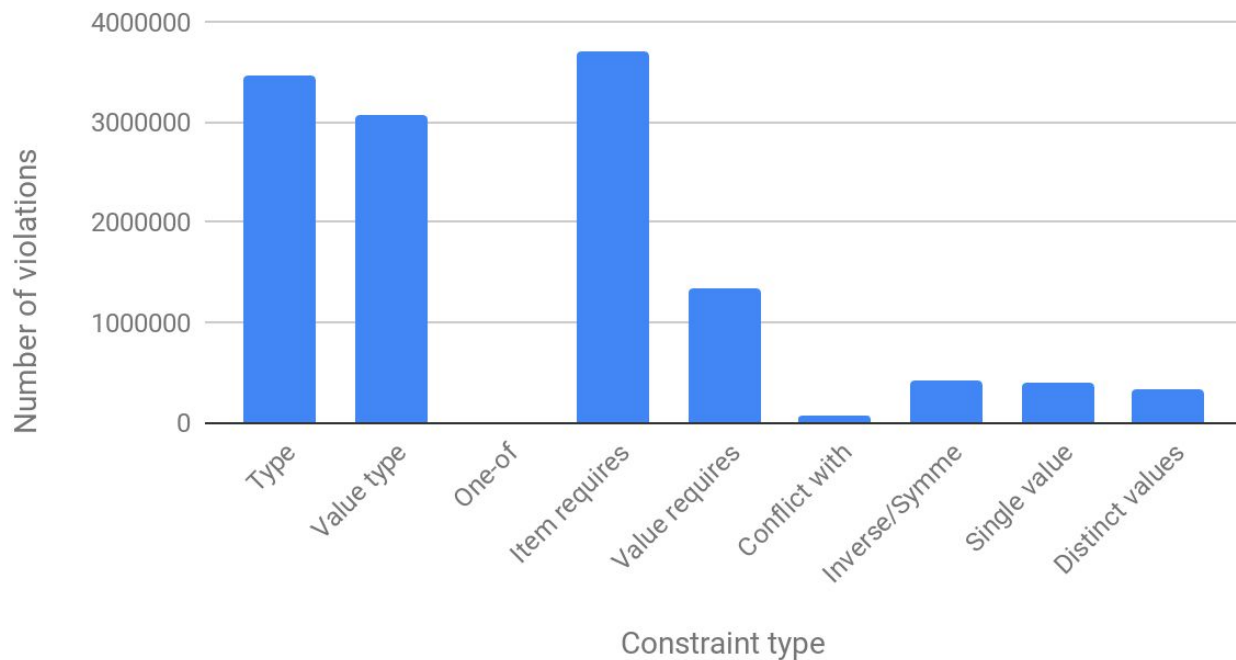
— Thomas Pellissier Tanon, Camille
Bourgau, Fabian Suchanek —

TELECOM
ParisTech



Knowledge bases are kind of messy

Wikidata constraints violations (July 2018)



KB \mathcal{K} with constraints

1. A-Box \mathcal{A}
2. T-Box \mathcal{T}
3. Constraints \mathcal{C}
 - consistency e.g. $\Gamma: \exists \text{gender}^- \sqsubseteq \{ \text{male, female, nonbinary} \}$
 - completeness e.g. $\Gamma: \exists \text{birthPlace} \sqsubseteq \exists \text{birthDate}$

Constraints

$\mathcal{K} = (\mathcal{A}, \mathcal{T})$ satisfies a constraint $\Gamma \in \mathcal{C}$ if $\mathcal{I}_{\mathcal{K}} \models \Gamma$
where $\mathcal{I}_{\mathcal{K}}$ is the canonical model of \mathcal{K}

Example:

If $\mathcal{A} = \{\text{Zeus} \text{ gender } \text{male}\}$
then \mathcal{K} satisfies

$\Gamma_1: \exists \text{gender} \sqsubseteq \{\text{male, female, nonbinary}\}$

Constraints as rules

Constraints could be written as UCQ rules

$$r_1: \exists \text{gender} \sqsubseteq \{ \text{male}, \text{female}, \text{nonbinary} \}$$
$$r_1(x): \exists y \text{gender}(y,x) \rightarrow x \in \{ \text{male}, \text{female}, \text{nonbinary} \}$$

Violations of $\Gamma(\vec{x})$ in \mathcal{K}

Minimal subset $\mathcal{V} \subseteq \mathcal{K}$ such that there exists \vec{a} such that \mathcal{V} violates $\Gamma(\vec{a})$ and \mathcal{K} violates $\Gamma(\vec{a})$

Example:

If $\mathcal{A} = \{\text{gender}(\text{Zeus}, \text{male}), \text{gender}(\text{Hera}, \text{woman})\}$
then $\mathcal{V} = \{\text{gender}(\text{Hera}, \text{woman})\}$ is a violation of
 $\Gamma_1: \exists \text{gender} \sqsubseteq \{\text{male}, \text{female}, \text{nonbinary}\}$

Atomic modifications

- Insertion: $+ \{ s2 \ p2 \ o2 \}$
- Deletion: $- \{ s1 \ p1 \ o1 \}$
- Modification: $- \{ s1 \ p1 \ o1 \} + \{ s2 \ p2 \ o2 \}$

Example:

$- \{ \text{Hera gender woman} \} + \{ \text{Hera gender female} \}$

Solution of a violation \mathcal{V} of $\Gamma(\vec{a})$ in \mathcal{K}

It is an atomic modification $(\mathcal{M}^+, \mathcal{M}^-)$ such that there exists $\mathcal{K}' \subseteq \mathcal{K}$ such that $(\mathcal{V} \cup \mathcal{K}' \cup \mathcal{M}^+) \setminus \mathcal{M}^-$ satisfies $\Gamma(\vec{a})$.

Example:

- $\{\text{Hera gender woman}\} + \{\text{Hera gender female}\}$ is a solution of $\mathcal{V} = \{\text{gender}(\text{Hera}, \text{woman})\}$

$\Gamma_1: \exists \text{gender}^- \sqsubseteq \{\text{male}, \text{female}, \text{nonbinary}\}$

Good solution

We want to make the KB
close to the real world

The edit history of the KB is
a provider of good solutions

The KB edit history provides past corrections

Before:

Matsuo Bashō (Q5676)...

place of birth Iga-Ueno Q3148112 ...

date of death

Revision history of "Iga-Ueno" (Q3148112)

View logs for this page (view abuse log)

Compare selected revisions

• (cur | prev) 16:39, 18 March 2019 Tpt (talk | contribs) .. (1,767 bytes) (+427) .. (Created claim: *instance of (P31): neighbourhood (Q123705)*) (undo) (Tag: PHP7)

• (cur | prev) 12:43, 3 July 2018 Dymitr (talk | contribs) .. (1,274 bytes) (+88) .. (Added [be-tarask] label: *Іга-Ўэна*) (undo | thank) (restore)

• (cur | prev) 08:37, 23 June 2018 Sian EJ (talk | contribs) .. (1,186 bytes) (+66) .. (Added [cy] label: *Iga-Ueno*) (undo | thank) (restore)

Compare selected revisions

Se Matsuo Bashō (Q5676)...

place of birth Iga-Ueno Q3148112 ...

0 references

Edit:

After:

Extracting past corrections

- Solving a violation, two options:

Matsuo Bashō (Q5676)...

place of birth

Iga-Ueno (Q3148112) ... [!]

Potential issues [X]

value type constraint [Help](#) [Discuss](#)

Values of *place of birth* statements should be instances of one of the following classes (or of one of their subclasses), but Iga-Ueno currently isn't:

- [geographical object](#)
- [fictional location](#)

date of death



Iga-Ueno (Q3148112)...

instance of

+

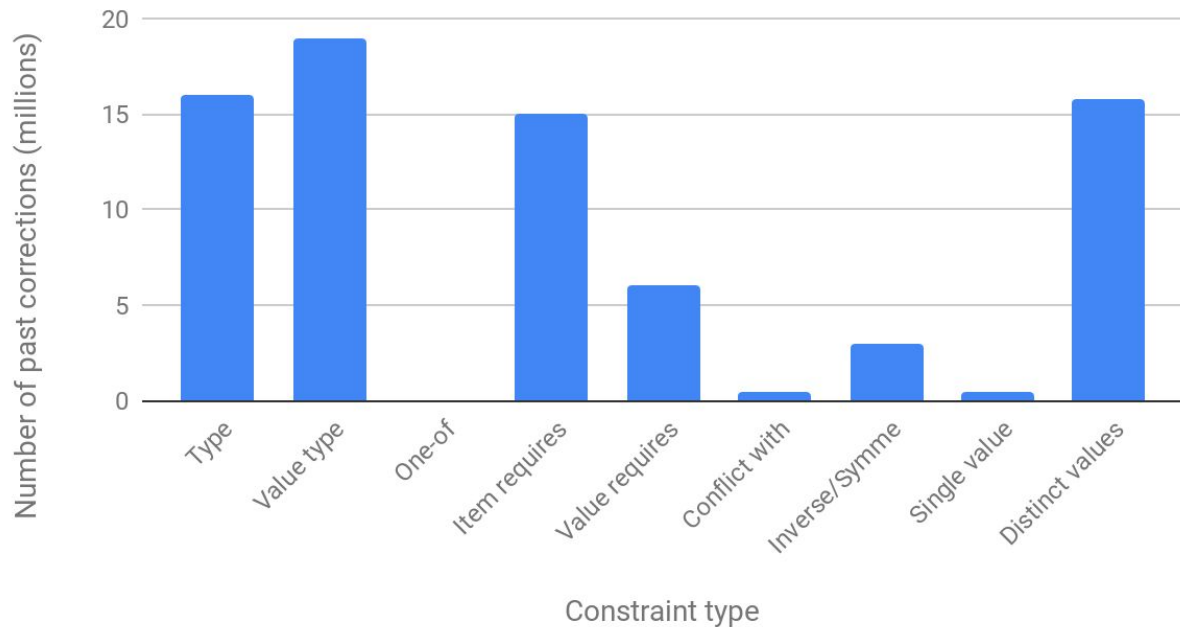
[geographical object](#)

▼ 0 references

- We look for such edits and check if they correct a violation

Applied on Wikidata

Extracted past corrections (July 2018)



rev 1223445: `placeOfBirth-valueType-violation(wd:MatsuoBashō, wd:Iga-Ueno)`

→ `+{ wd:Iga-Ueno wdt:type wd:geoObject }`

rev 2334569: `gender-oneOf-violation(wd:Nefertiti, wd:woman)`

→ `- { wd:Nefertiti wdt:gender wd:woman } +{ wd:Nefertiti wdt:gender wd:female }`

There are patterns for finding the good solutions

Nefertiti (Q40930)...

sex or gender woman Q467 ...

date of birth

Issues ✕
one-of constraint Help Discuss
The value for sex or gender should be one of the following:

- male
- female
- intersex
- hermanbrodite



Nefertiti (Q40930)...

sex or gender female

0 references

$[I_1(?s, wd:woman)]: \rightarrow - \{ ?s \ wdt:gender \ wd:woman \}$

$+ \{ ?s \ wdt:gender \ wd:female \}$

Why rules?

- Explainable
- Works well with new entities

Mining correction rules

- **Start** from the past corrections
- **Generalize** using the KB state at the correction

Using **AMIE** (slightly modified)
and standard confidence

Wikidata evaluation

178k rules mined on 80% of the past corrections

Some top rules:

Single Value:

[gender-singleValue-violation(?s, wd:maleOrganism)] ?s wdt:sportsTeam ?t
→ - { ?s wdt:gender wd:maleOrganism }

One-of :

[mannerOfDeath-oneOf-violation(?s wd:trafficAccident)]
→ - { ?s wdt:mannerOfDeath wd:trafficAccident }
+ { ?s wdt:causeOfDeath wd:trafficAccident }

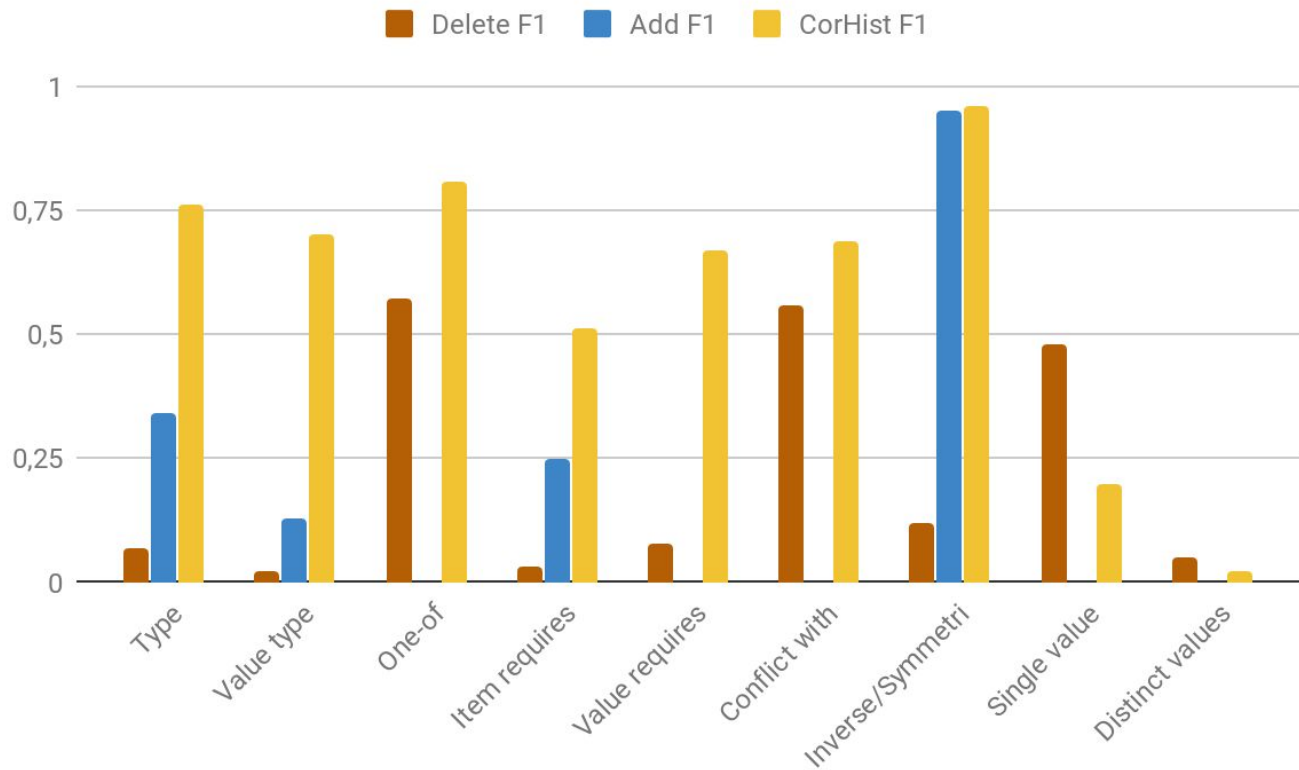
Evaluation on the past corrections (on Wikidata)

- 178k rules mined on 80% of the past corrections
- Apply the rules on the other 20% known corrections
- Compute precision and recall
- Baselines
 - Remove the violation
 - Add the missing triple (if possible)

Wikidata evaluation: Some results

Micro averages

$$F1 = \frac{2 \times \textit{precision} \times \textit{recall}}{\textit{precision} + \textit{recall}}$$



User evaluation: suggest corrections to Wikidata

- Experiment on three months
- 47 participants
- 50k suggested corrections

Francesco Belinzeri [Q57082102]

Auto | it

Francesco Belinzeri is a [Italian sculptor, painter, and architect](#).

Violation

An entity should not have a statement for [country of citizenship](#) if it also has a statement for [sex or gender](#) with value [male non-human organism](#).

Possible correction

Edit [statement \(Q57082102, sex or gender, male non-human organism\)](#). Setting value to: [male](#)

User evaluation results

- **Inverse/Symmetric**: 22k actions, 92% approval
- **Value requires statement** and **Conflicts with**: 1k actions each, 80% approval
- **Others**: between 30 et 700 actions, approval between 20% and 50%

Biased by what has been done (or not) by bots

- Some huge easy completions
- Mostly hard stuff remains

Contribution

- Introduction and formalization of the problem of learning corrections from a KB history
- A competitive rule mining approach

Future work

- Interesting problems
 - *There are two birth places*
 - *A birth date is missing*

- Applications
 - Suggest edits
 - Fight vandalism

Thank you!

Paper: <https://thomas.pellissier-tanon.fr/papers/2019-WWW-corhist.pdf>

Game: <https://tools.wmflabs.org/wikidata-game/distributed/#game=43>

Dataset: <https://doi.org/10.6084/m9.figshare.7712720>

Code: <https://github.com/Tpt/corhist>

Wikidata History SPARQL endpoint: <https://wdhqs.wmflabs.org>