# 10 steps to integrate CIViCdb with other public data in Wikidata

Elvira Mitraka (1), Andra Waagmeester (2), Núria Queralt-Rosinach (3), Sebastian Burgstaller-Muehlbacher (3), Gregory Stupp (3) Lynn Schriml (1), Josh F. McMichael (4), Benjamin Ainscough (4), Malachi Griffith (4), Obi L. Griffith (4), Andrew I. Su (3) and Benjamin M. Good (3)

(1) University of Maryland Baltimore, Baltimore, MD, USA,(2) Micelio, Antwerp, Belgium, (3) Department of Integrative Structural and Computational Biology, the Scripps Research Institute, La Jolla, USA, (4) McDonnell Genome Institute, Department of Medicine, Washington University School of Medicine, St Louis, MO USA

Step 1: Select a public license

https://creativecommons.org/publicdomain/zero/1.0/
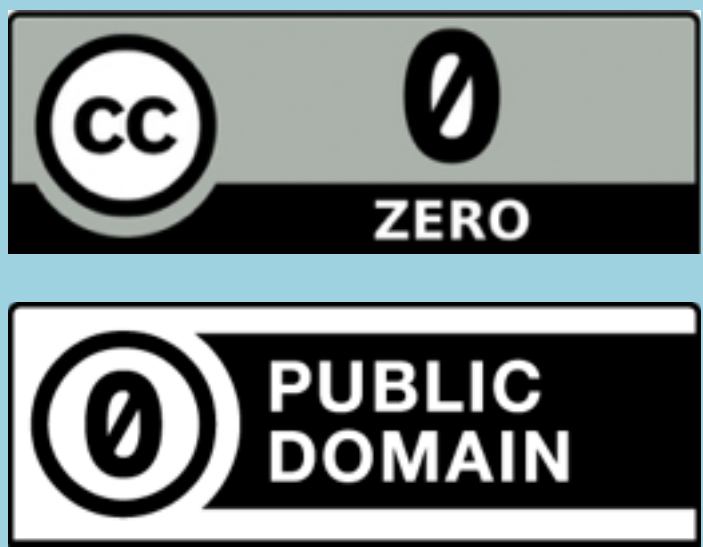
Step 2: Be(come) FAIR & Notable

FAIR data

Notable items fulfil at least 1 of the following criteria:

1. Contains at least one valid sitelink
2. It refers to an instance of a clearly identifiable conceptual or material entity
3. It fulfils some structural need

FAIR: https://www.force11.org/group/fairgroup
Notability https://www.wikidata.org/wiki/Wikidata:Notability
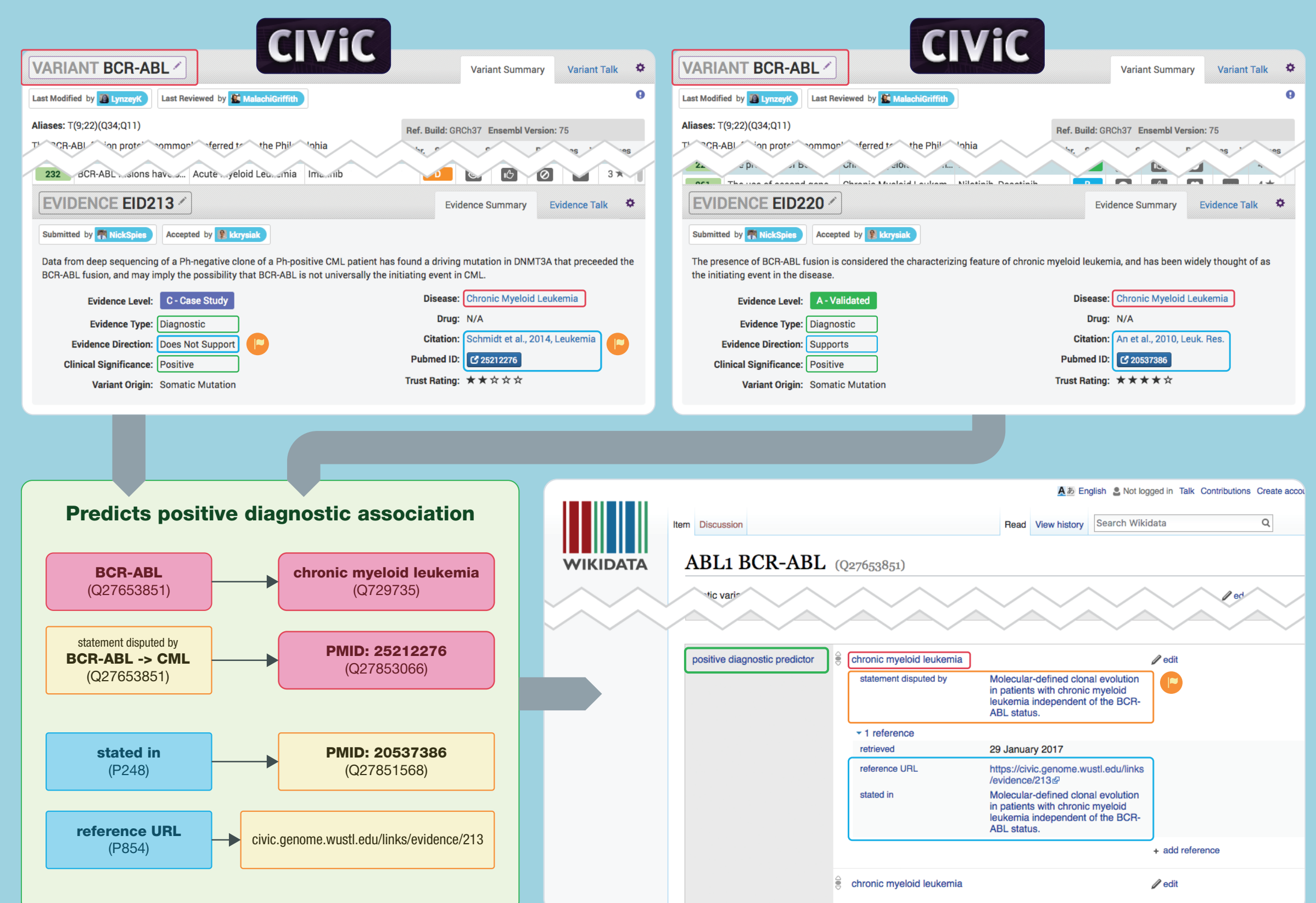
Step 3: Join Wikidata communities

Wikiprojects:
- https://www.wikidata.org/wiki/Wikidata:WikiProject_Molecular_biology
- https://www.wikidata.org/wiki/Wikidata:WikiProject_Medicine
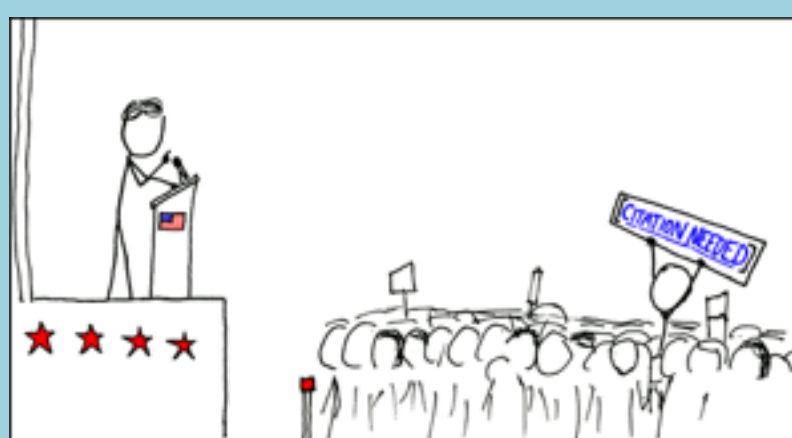- https://www.wikidata.org/wiki/Wikidata:WikiProject_Ontology

Step 4: Reach consensus on underlying data model

Data and Knowledge Production

Millions of raw sequence reads are produced for a patient tumor.

Sequences are aligned to the reference genome and tumor-specific events predicted.

Data are reviewed and validation experiments performed to identify high quality events.

Events are annotated and scored in an effort to predict events of functional significance.

CIViC Curation

Crowdsourced curation efforts, moderated by experts in oncology and bioinformatics, help to build a knowledge-base of clinical interpretations of variants in cancer, describing the therapeutic, prognostic, diagnostic, and predisposing relevance of inherited and somatic variants of all types. Anyone may sign up to be a curator, add evidence, suggest changes to records, and discuss ongoing curation efforts.
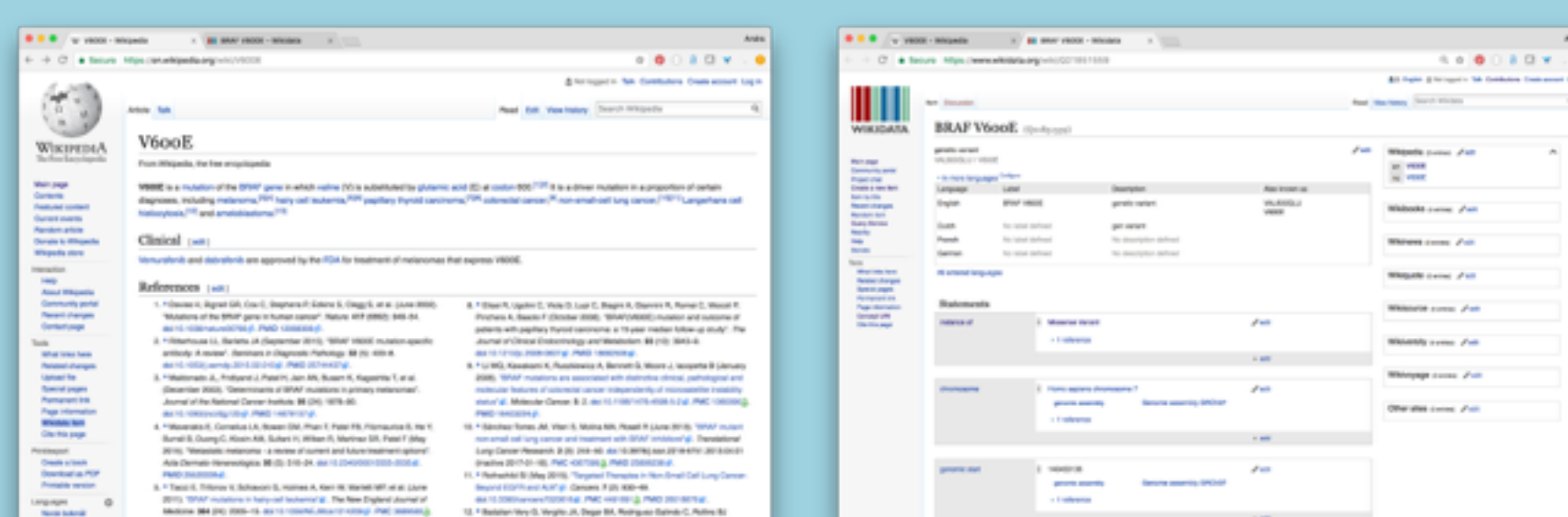
Step 5: Show evidence

WikiCite aims at building a large bibliographic database in Wikidata to serve all Wikimedia projects. Citations can now be added::
Source Metadata:
https://tools.wmflabs.org/sourcemd/
PMIDTool:
https://www.wikidata.org/wiki/Wikidata:WikiProject_Source_MetaData/PMIDTool

wikicite: https://meta.wikimedia.org/wiki/WikiCite
Evidence patterns: https://www.wikidata.org/wiki/User:ProteinBoxBot/evidence

Step 6: Create a canonical example

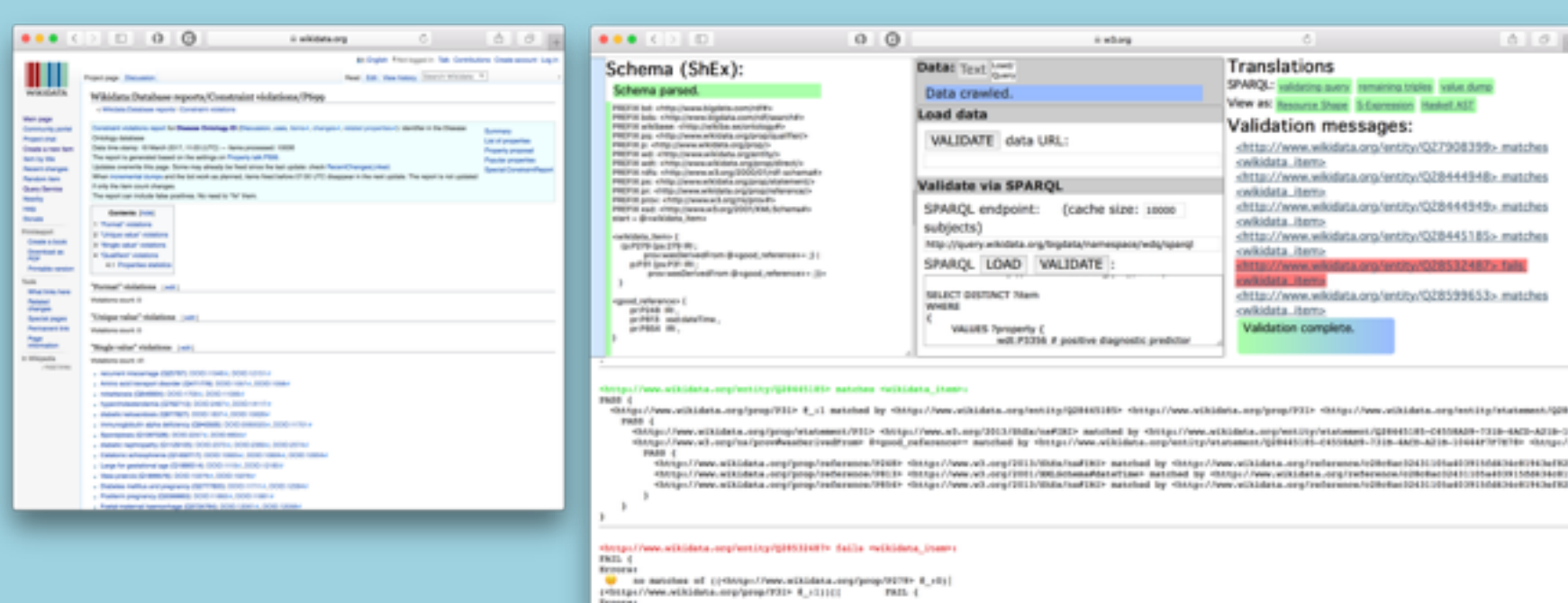Wikipedia: https://en.wikipedia.org/wiki/V600E
Wikidata: https://www.wikidata.org/wiki/Q21851559

Step 7: Automate

- WikidataIntegrator is a python library for reading and writing to Wikidata/Wikibase.
- It interacts with the Wikidata API
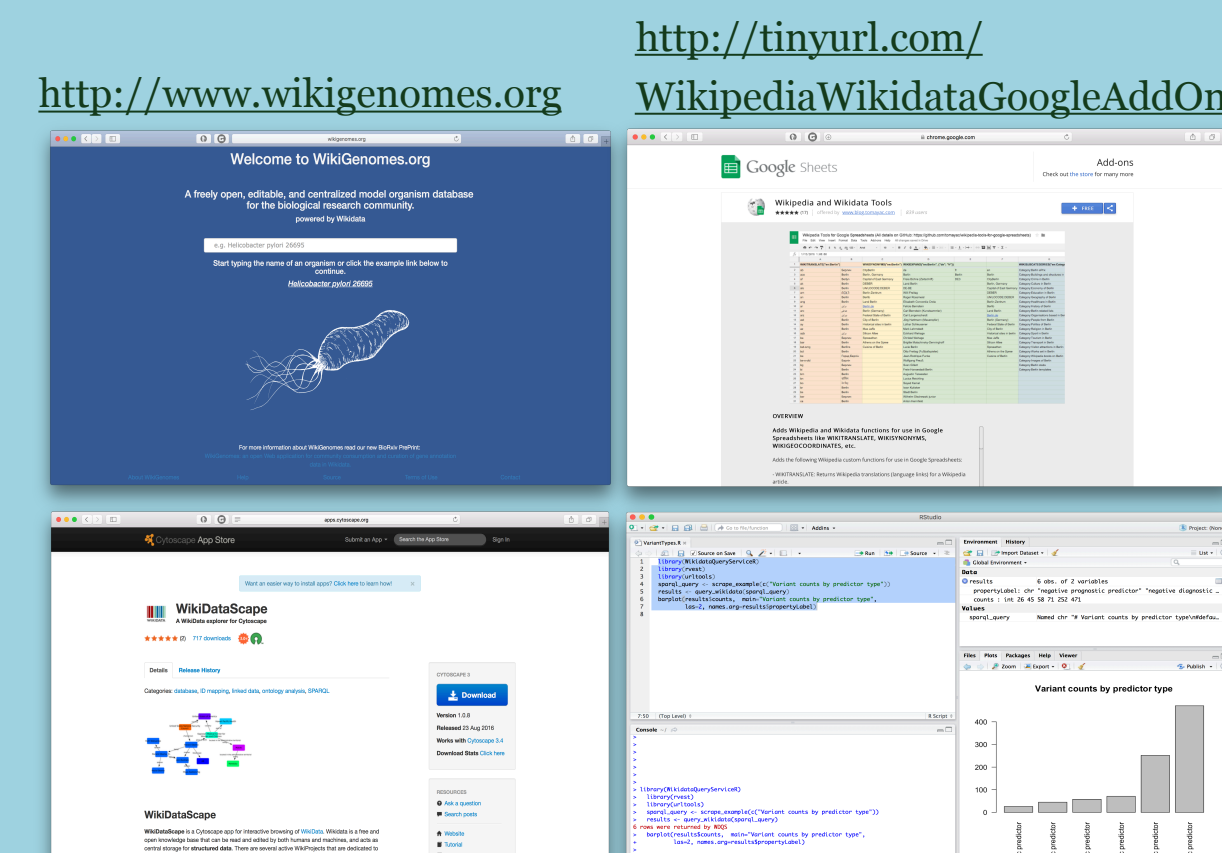- It uses SPARQL queries to ensure proper concept resolution, which prevent creation of duplicate items
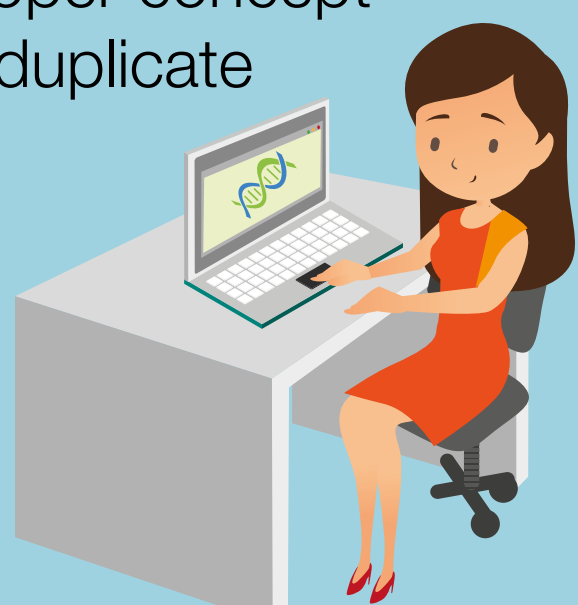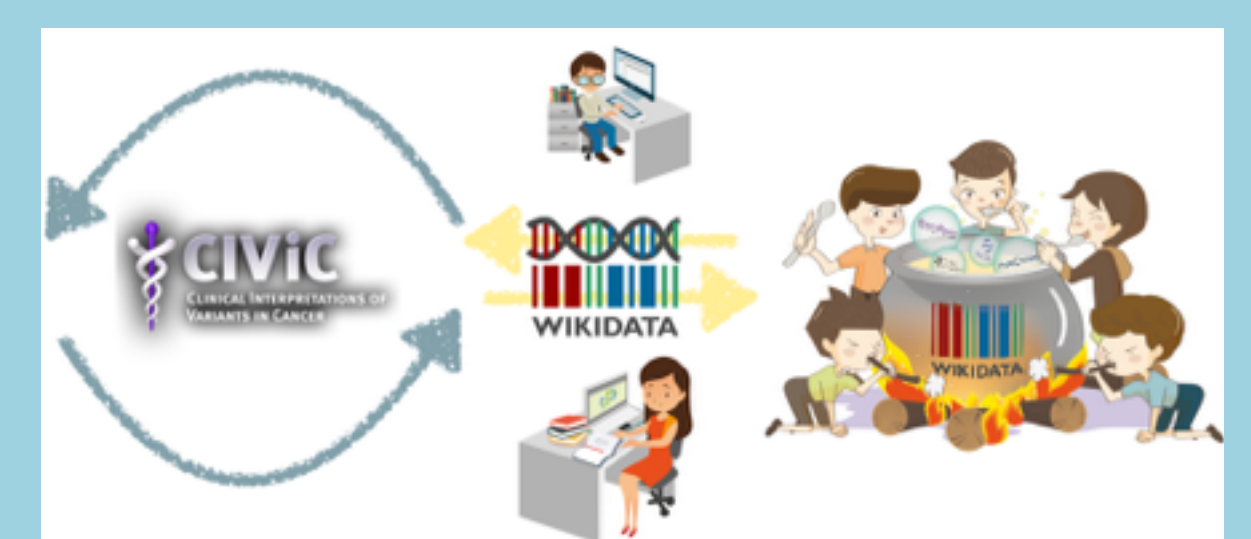
Step 8: Validate

Wikidata constraint violations:
https://www.wikidata.org/wiki/Wikidata:Database_reports/Constraint_violations
Shape expressions: http://tinyurl.com/wikidataShExDemo

Step 9: Reuse & Tooling

http://www.wikigenomes.org
http://tinyurl.com/WikipediaWikidataGoogleAddOn

http://apps.cytoscape.org/apps/wikidatascape
https://cran.r-project.org/web/packages/WikidataQueryServiceR/index.html

Step 10: Update

An open source automation server is used to regularly run all gene wiki bots to synchronise between original sources an Wikidata projects. https://www.wikidata.org/wiki/User:ProteinBoxBot/Automation

Jenkins: http://jenkins.io
PAWS: https://www.mediawiki.org/wiki/PAWS

Acknowledgements