Literature Survey

Study the role of Wikidata in enhancing the productivity of the Wikipedia writers and enhance Wikidata for Hindi and Telugu using gamification

Name - Himanshu Maheshwari **Mentors -** Tushar Abhishek and Nikhil Pattiisapu
Roll No - 20171033

Abstract

Wikidata is a free structured knowledge base that can be read and edited by both humans and Machines. Wikipedia is a multilingual, web-based, free-content encyclopedia project supported by the Wikimedia Foundation and based on a model of openly editable content. Wikidata has the potential to provide with factual data that could be leveraged by the Wikipedia writers.

Gamification is the application of typical elements of game playing (e.g. point scoring, competition with others, rules of play) to other areas of activity (here enhancing Wikidata for Hindi and Telugu), to encourage engagement with a product or service. This survey talks about a few games that aims to enhance *english* wikidata viz. The Wikidata Game, Wikidata Game abc etc.

Wikidata

Wikidata is a free and open knowledge base that can be read and edited by both humans and machines. The Wikidata repository consists mainly of items, each one having a label, a description and any number of aliases. In Wikidata, items are used to represent all the *things* in human knowledge, including topics, concepts, and objects. For example, the "1988 Summer Olympics", "love", "Elvis Presley", and "gorilla" are all items in Wikidata. It could be used to store images, geo-location, sound etc. which could be directly used in Wikipedia articles.

Wikidata is the centralized, linked data repository for all Wikimedia projects. This means that all Wikimedia projects (Commons and Wikipedia for instance) can pull the information from the same central place. This also means that all 300+ language versions of Wikipedia can pull data from Wikidata as well. There is incredible potential for more access to information, more consistency across different languages, and the ability for any language-speaker to contribute more equitably. Beyond the effect it is having in Wiki-verse, Wikidata is machine readable. This means that digital assistants, AI, bots, and scripts can interact with Wikidata’s structured, linked data. It has evolved from an experimental semantic web database to an inter-language linking hub for Wikipedia articles, to now being an engine for capturing relationships among numerical, text, visual and graphical content.

There are two possible Wikipedia writers - Humans and Bots. Wikidata provides data to Wikipedia. Its members are ingesting, cleansing and improving data; discussing data modelling issues; contacting new data providers. All this effort leads to mammoth amounts of data being collected.(just to give idea Wikidata has more than 18 million scientific publications indexed!) Equipped with such a large quantity of interlinked data (Wikidata is free and its data could be use without copyright restrictions) Wikipedia writers could easily create wiki pages. Thus human writers could use Wikidata as a knowledge pool to take data from.

However the main role that wikidata plays in comes from its machine friendly nature. This machine friendly nature could be leveraged by bots to create wikipedia pages. (Do note that unlike wikidata, knowledge base like DBPedia etc. uses wikipedia to enrich their knowledge base). [6], [7] & [10] talks about how data from Wikidata can be directly displayed on Wikimedia projects using bots. Following is a brief summary of what these article discusses.

It is possible for infoboxes on Wikipedia to use data directly from Wikidata. There are several examples of Wikidata fed infoboxes on Wikipedia, these include:

- World Heritage sites on English Wikipedia e.g Giza pyramid complex.
- Telescopes on English Wikipedia e.g Telescope Array Project.

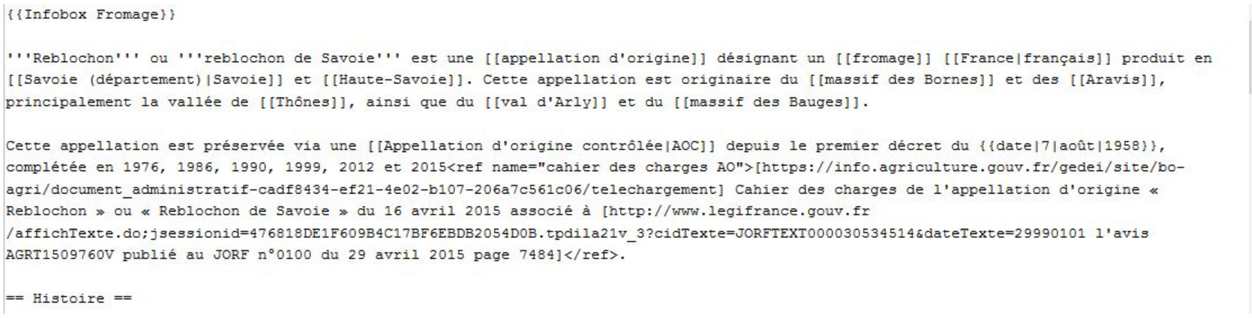
The French Wikipedia articles about monuments and artists have adopted more than 40,000 Wikidata-driven infoboxes. Such infoboxes and lists are an immensely useful tool to eliminate the duplication of efforts across Wikipedia’s different language communities, as the data can be curated in a central database. At the same time, they have a large impact in terms of visibility of the data thanks to the prominence of Wikipedia on the Internet. Some of the smaller (or medium-sized) language communities, such as the Catalan Wikipedia, have been pioneers in this area, creating infobox templates with interactive maps, detailed career information for biographies and various other information directly fetched from Wikidata. In fact, in fall 2017 the Catalan community announced that over 50% of the 550’000 articles on Catalan Wikipedia were using data from Wikidata.

In addition to the direct inclusion of data from Wikidata in Wikipedia articles, Wikidata has also been used as a basis for the automatic generation of article stubs that could then be extended by Wikipedia editors. This approach is especially useful for harmonizing the structure of articles in specific areas. One of the tools for this is the “Mbabel Article Generator”, which was first developed for a New-York-based project, about museums, and was then extended by a Brazilian team to other classes of subjects, such as works of art, books, films, earthquakes, and journals. Here again, a Wikidata-based tool can help reduce the workload of Wikipedians by using centrally curated data that can be applied across language communities. Another tool that serves a similar purpose for biographical articles is “PrepBio”.

[11] talks about Mbabel tool. The purpose of the tool is to facilitate the process of creating entries for Wikipedia by providing a simplified outline, automatically generated from data present in another Wikimedia sister project, Wikidata.

With the parser function or Lua code, it is possible to display labels, descriptions, values, references, and a lot of other information stored on Wikidata. The two main client functionalities, parser function and access via Lua, can be enabled together on the wikis. Data can also be accessed with Lua modules, which are much more flexible. If your wiki does not contain a module you can copy it from another wiki and add documentation. SPARQL could also be used to access the data. Some templates use modules to access Wikidata data. They are as simple to use as regular templates.

Below image contain A simple code that can be used on French Wikipedia to build an infobox about cheese.



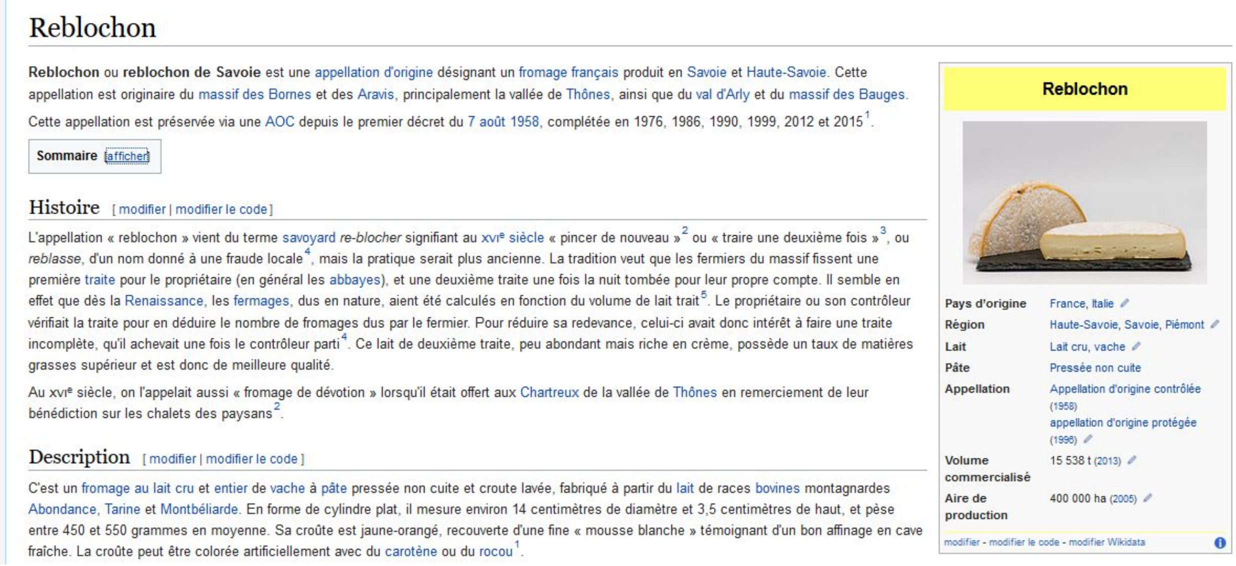
```
{{Infobox Fromage}}

'''Reblochon''' ou '''reblochon de Savoie''' est une [[appellation d'origine]] désignant un [[fromage]] [[France|français]] produit en
[[Savoie (département)|Savoie]] et [[Haute-Savoie]]. Cette appellation est originaire du [[massif des Bornes]] et des [[Aravis]],
principalement la vallée de [[Thônes]], ainsi que du [[val d'Arly]] et du [[massif des Bauges]].

Cette appellation est préservée via une [[Appellation d'origine contrôlée|AOC]] depuis le premier décret du {{date|7|août|1958}},
complétée en 1976, 1986, 1990, 1999, 2012 et 2015<ref name="cahier des charges AO">[https://info.agriculture.gouv.fr/gedei/site/bo-
agri/document_administratif-cadf8434-ef21-4e02-b107-206a7c561c06/telechargement] Cahier des charges de l'appellation d'origine «
Reblochon » ou « Reblochon de Savoie » du 16 avril 2015 associé à [http://www.legifrance.gouv.fr
/affichTexte.do;jsessionid=476818DE1F609B4C17BF6EBDB2054D0B.tpdila21v_3?cidTexte=JORFTEXT000030534514&dateTexte=29990101 l'avis
AGRT1509760V publié au JORF n°0100 du 29 avril 2015 page 7484]</ref>.

== Histoire ==
```

The infobox fills itself using data from the Wikidata item linked to the article.



Reblochon

Reblochon ou **reblochon de Savoie** est une appellation d'origine désignant un fromage français produit en Savoie et Haute-Savoie. Cette appellation est originaire du massif des Bornes et des Aravis, principalement la vallée de Thônes, ainsi que du val d'Arly et du massif des Bauges.

Cette appellation est préservée via une AOC depuis le premier décret du 7 août 1958, complétée en 1976, 1986, 1990, 1999, 2012 et 2015[1].

Histoire [modifier | modifier le code]

L'appellation « reblochon » vient du terme savoyard *re-blocher* signifiant au xvie siècle « pincer de nouveau »[2] ou « traire une deuxième fois »[3], ou *reblasse*, d'un nom donné à une fraude locale[4], mais la pratique serait plus ancienne. La tradition veut que les fermiers du massif fissent une première traite pour le propriétaire (en général les abbayes), et une deuxième traite une fois la nuit tombée pour leur propre compte. Il semble en effet que dès la Renaissance, les fermages, dus en nature, aient été calculés en fonction du volume de lait trait[5]. Le propriétaire ou son contrôleur vérifiait la traite pour en déduire le nombre de fromages dus par le fermier. Pour réduire sa redevance, celui-ci avait donc intérêt à faire une traite incomplète, qu'il achevait une fois le contrôleur parti[4]. Ce lait de deuxième traite, peu abondant mais riche en crème, possède un taux de matières grasses supérieur et est donc de meilleure qualité.

Au xvie siècle, on l'appelait aussi « fromage de dévotion » lorsqu'il était offert aux Chartreux de la vallée de Thônes en remerciement de leur bénédiction sur les chalets des paysans[2].

Description [modifier | modifier le code]

C'est un fromage au lait cru et entier de vache à pâte pressée non cuite et croute lavée, fabriqué à partir du lait de races bovines montagnardes Abondance, Tarine et Montbéliarde. En forme de cylindre plat, il mesure environ 14 centimètres de diamètre et 3,5 centimètres de haut, et pèse entre 450 et 550 grammes en moyenne. Sa croûte est jaune-orangé, recouverte d'une fine « mousse blanche » témoignant d'un bon affinage en cave fraîche. La croûte peut être colorée artificiellement avec du carotène ou du rocou[1].

Reblochon	
Pays d'origine	France, Italie
Région	Haute-Savoie, Savoie, Piémont
Lait	Lait cru, vache
Pâte	Pressée non cuite
Appellation	Appellation d'origine contrôlée (1958) appellation d'origine protégée (1996)
Volume commercialisé	15 538 t (2013)
Aire de production	400 000 ha (2005)

Commons creator template [8] uses Wikidata with arbitrary access to provide information about the creators of the works. All [6], [7], [8], [10] and [11] talks about using Wikidata for Wikimedia projects especially Wikipedia using simple bots.

[12] talks about about a potential extension to Wikidata that would allow for automated generation of articles on Wikipedias from Wikidata data.

Data from Wikidata is available in RDF dumps. These RDF triplets could be used for automatic text generation. These text could be used in Wikipedia article generation. RDF to text is an active research problem. There are various research paper that provides solution to this problem [13] is one such research paper.

Thus to conclude there are two kinds of writers that contribute to Wikipedia - Humans and Bots. Wikidata provides large amount of data to Humans. Its machine friendly nature is leveraged by bots for automatic text generation. Currently a lot of info-boxes are created automatically using data from Wikidata.

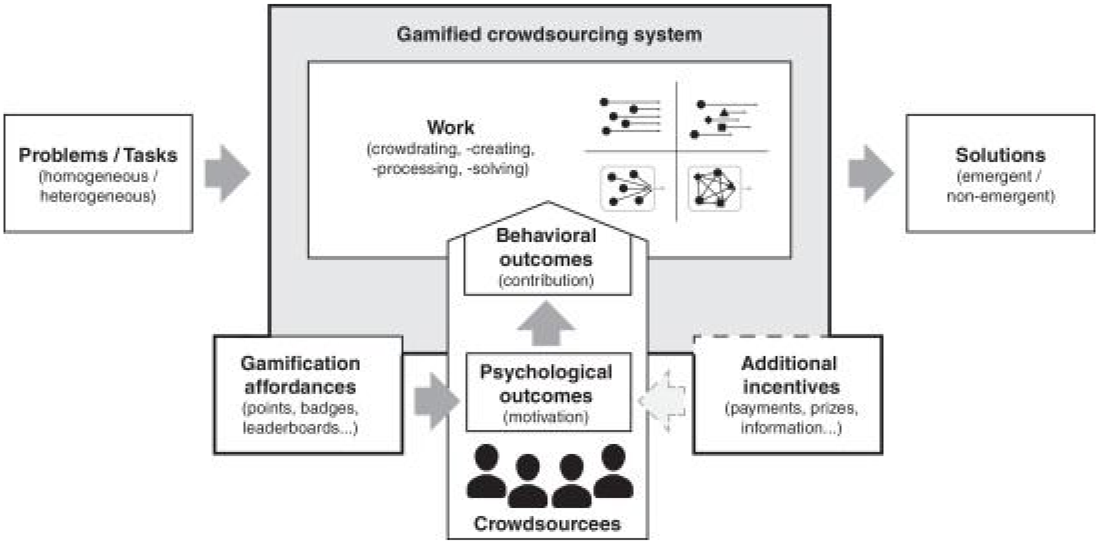
Before talking about enhancing Wikidata for Hindi and Telugu using gamification it is imperative to talk about how Wikidata is created. Wikidata gets its data through crowdsourcing. Crowdsourcing is the practice of obtaining information or input into a task or project by enlisting the services of a large number of people, either paid or unpaid, typically via the Internet.

Gamification can be defined as the application of game-design elements and game principles in non-game contexts. Gamification commonly employs game design elements to improve user engagement, organizational productivity, flow, learning, crowdsourcing, employee recruitment and evaluation, ease of use, usefulness of systems, physical exercise, traffic violations, voter apathy, and more.

At first we will talk about the role of gamification in crowdsourcing in general and then in context of Wikidata. Then we will talk about some already existing games aimed at enhancing Wikidata.

Crowdsourcing through Gamification

Crowdsourcing systems are increasingly gamified, that is, organizations seek to make the crowdsourced work activity more like playing a game in order to provide other motives for working than just monetary compensation. These gamified system increases participant interest into the task providing much better results..Such gamified crowdsourcing systems are increasing, and are a major application area of gamification. the results of various reviews [15] indicate that gamification has been an effective approach for increasing crowdsourcing participation.


Gamified crowdsourcing system
Problems / Tasks
(homogeneous / heterogeneous)
Work
(crowdrating, -creating, -processing, -solving)
Solutions
(emergent / non-emergent)
Behavioral outcomes
(contribution)
Gamification affordances
(points, badges, leaderboards...)
Psychological outcomes
(motivation)
Additional incentives
(payments, prizes, information...)
Crowdsourcees

Conceptual Framework of Gamified Crowdsourcing Systems.

Google Image Labeler was a rather early example of a crowd sourced game. It was launched back in 2006 when Google was seeking to improve the accuracy of its image database. They wanted to make sure that the images which came up during searches on Google Images were the most relevant to user queries.

To achieve this massive undertaking, the company decided to integrate this task into a game whose structure was based on the original ESP game (an idea in computer science for addressing the problem of creating difficult metadata- effectively the original concept behind crowdsourcing).

Participants were each paired with a partner online. Each was shown an image and asked to generate as many labels as possible. The partners were awarded points when their labels matched.

This game is largely driven by social influence, even though the players were connected to just one other person (i.e. partner). The success of Google Image Labeler paved way for future Gamified crowdsourcing systems.

Existing Wikidata Games

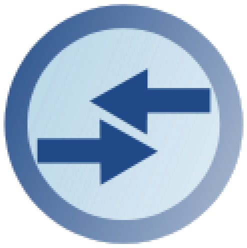
A lot of games exists to enhance Wikidata, however almost all of them focuses on English wikidata and as such there are no existing games for Hindi or Telugu Wikidata.

A lot of games are hosted on Wikidata Game and Wikidata - The Distributed Game ([17] & [18]). These games are basic questions and answer games. It was created by Magnus Manske. Both of these are similar in their approach.
They can perform up to three actions on every decision the gamer takes:

- Store the decision centrally, for Recent Changes, user contributions, statistics etc.
- Feed the decision back to the remote API that provided the game tile, if only to mark this tile as “done” and not present it again.
- Perform one or more edits on Wikidata.

Some examples of the games that are hosted are:

1. **Merge items**
   Some topics have duplicate items on Wikidata. Two items with the same title or alias will be suggested to you. The aim of this game is to merge identical topics, tag items as different, or skip an item pair if you are not sure.

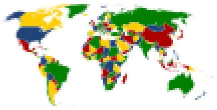
2. **Person**
   Many items about people on Wikidata have no "instance of" property. The aim of the game is to decide if one of them is a human, or skip the item if you are not sure.

3. **Gender**
   Many items about people on Wikidata have no gender property set. The aim of the game is to add a sex/gender statement "male" or "female", or skip the item if you are not sure.

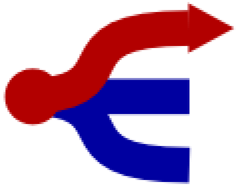
4. **Occupation**
   The people shown in the game have no occupation added, but their articles suggest something else. Thus the aim of the game is to provide the occupations of these people.

5. **Alma mater**
   The people shown in the game have no alma mater added, but their articles suggest something else. Thus the aim of the game is to provide the occupations of these people.

6. **Country of citizenship**
   The people shown have a birth place, but no country of citizenship added. Thus the aim of the game is to provide the citizenship of these people.

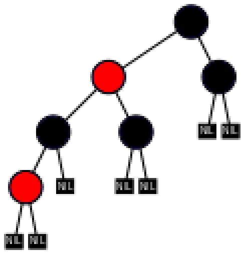


7. **Disambiguation items**
   The items shown could represent disambiguation pages, and the aim of the game is to be marked as such!

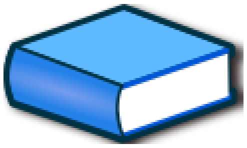


8. **Date**
   The people shown have missing date of birth or death. The aim of the game it provide this information.
9. **Image**
   The items shown have no image in their wikidata entry, but their Wikipedia articles do. The aim of the game is to provide this information.

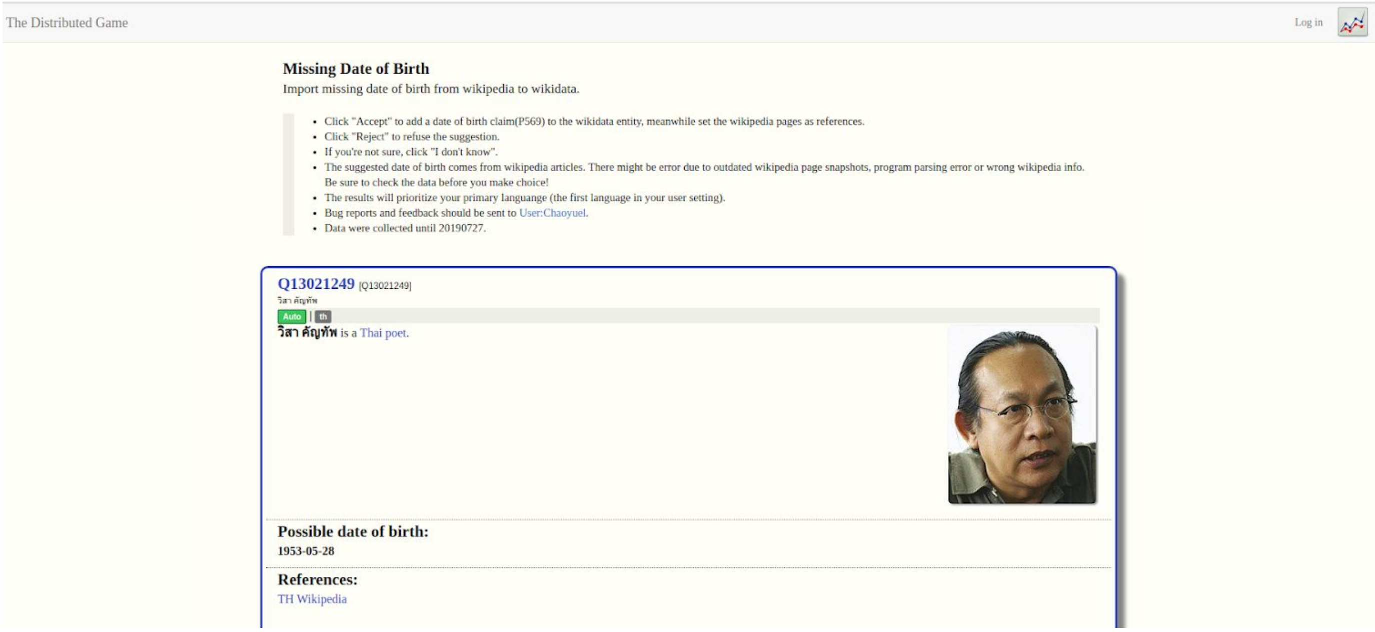
10. **Commons category**
    The items shown have no Commons category, but one with the exact same name exists. The aim of the game is to provide this commons category.



11. **Books without author**
    The books shown have no author on Wikidata, but their Wikipedia articles might. The aim of the game is to provide the author information.



Below is a screenshot of one such game: As we can see that it talks about a Thai poet and crowdsources his date of birth.



Missing Date of Birth
Import missing date of birth from wikipedia to wikidata.

- Click "Accept" to add a date of birth claim(P569) to the wikidata entity, meanwhile set the wikipedia pages as references.
- Click "Reject" to refuse the suggestion.
- If you're not sure, click "I don't know".
- The suggested date of birth comes from wikipedia articles. There might be error due to outdated wikipedia page snapshots, program parsing error or wrong wikipedia info. Be sure to check the data before you make choice!
- The results will prioritize your primary languange (the first language in your user setting).
- Bug reports and feedback should be sent to User:Chaoyuel.
- Data were collected until 20190727.

Q13021249 [Q13021249]
วิสา คัญทัพ

วิสา คัญทัพ is a Thai poet.

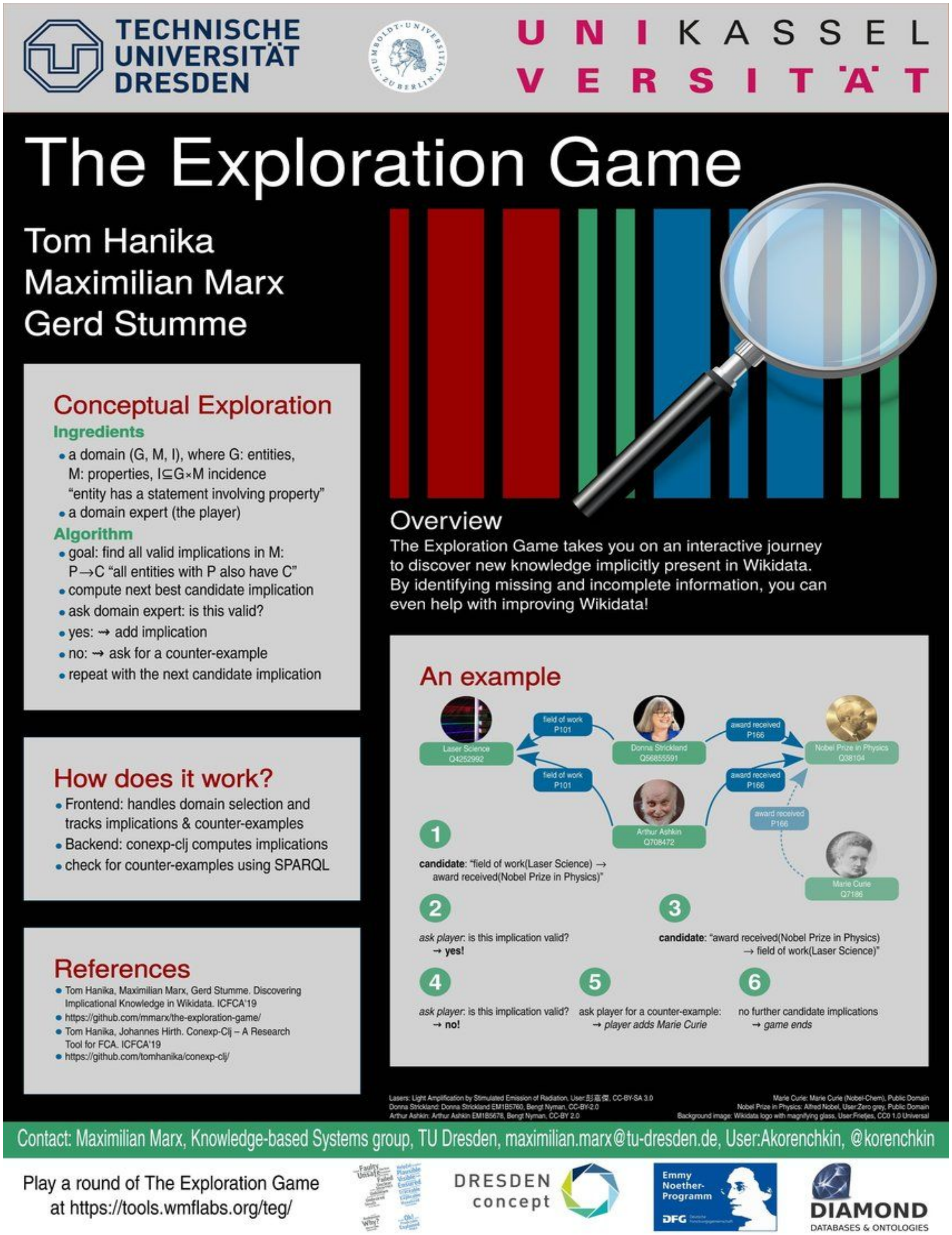
Possible date of birth:
1953-05-28

References:
TH Wikipedia

However do note that these games are not actually games in traditional senses. They are basically question and answer without much gaming principles involved.

Another example of gaming in crowdsourcing wikidata is - **The Exploration Game.**
The Exploration Game ([19]) is an interactive game for exploring implicit knowledge in Wikidata, based on the theory laid out in the publication "Discovering Implicational Knowledge in Wikidata" and

the Conceptual Exploration techniques from Formal Concept Analysis. The following poster describes the prototype implementation of the game. The development on this tool continues. By identifying missing and incomplete information gamer could also help in improving Wikidata. Though the aim of the game is not enhance Wikidata but it serves as a side purpose.



TECHNISCHE UNIVERSITÄT DRESDEN

UNIKASSEL VERSITÄT

The Exploration Game

Tom Hanika
Maximilian Marx
Gerd Stumme

Conceptual Exploration

Ingredients

- a domain (G, M, I), where G: entities, M: properties, I⊆G×M incidence "entity has a statement involving property"
- a domain expert (the player)

Algorithm

- goal: find all valid implications in M: P→C "all entities with P also have C"
- compute next best candidate implication
- ask domain expert: is this valid?
- yes: → add implication
- no: → ask for a counter-example
- repeat with the next candidate implication

Overview

The Exploration Game takes you on an interactive journey to discover new knowledge implicitly present in Wikidata. By identifying missing and incomplete information, you can even help with improving Wikidata!

How does it work?

- Frontend: handles domain selection and tracks implications & counter-examples
- Backend: conexp-clj computes implications
- check for counter-examples using SPARQL

An example

1 **candidate**: "field of work(Laser Science) → award received(Nobel Prize in Physics)"

2 *ask player*: is this implication valid? → **yes!**

3 **candidate**: "award received(Nobel Prize in Physics) → field of work(Laser Science)"

4 *ask player*: is this implication valid? → **no!**

5 ask player for a counter-example: → *player adds Marie Curie*

6 no further candidate implications → *game ends*

References

- Tom Hanika, Maximilian Marx, Gerd Stumme. Discovering Implicational Knowledge in Wikidata. ICFCA'19
- https://github.com/mmarx/the-exploration-game/
- Tom Hanika, Johannes Hirth. Conexp-Clj – A Research Tool for FCA. ICFCA'19
- https://github.com/tomhanika/conexp-clj/

Lasers: Light Amplification by Stimulated Emission of Radiation, User:彭嘉傑, CC-BY-SA 3.0
Donna Strickland: Donna Strickland EM1B5760, Bengt Nyman, CC-BY-2.0
Arthur Ashkin: Arthur Ashkin EM1B5678, Bengt Nyman, CC-BY 2.0
Marie Curie: Marie Curie (Nobel-Chem), Public Domain
Nobel Prize in Physics: Alfred Nobel, User:Zero grey, Public Domain
Background image: Wikidata logo with magnifying glass, User:Frietjes, CC0 1.0 Universal

Contact: Maximilian Marx, Knowledge-based Systems group, TU Dresden, maximilian.marx@tu-dresden.de, User:Akorenchkin, @korenchkin

Play a round of The Exploration Game at https://tools.wmflabs.org/teg/

DRESDEN concept

Emmy Noether-Programm DFG

DIAMOND DATABASES & ONTOLOGIES

❒❒❒

References

1. https://scholarworks.iupui.edu/bitstream/handle/1805/16690/Lemus-Rojas_Pintscher_Wikidata_2017-07-03.pdf
2. https://en.wikipedia.org/wiki/Wikipedia:About
3. https://tools.wmflabs.org/admin/tool/wikidata-game
4. https://en.wikipedia.org/wiki/Wikidata
5. https://wikiedu.org/blog/2019/06/03/why-is-wikidata-important-to-you/
6. https://www.wikidata.org/wiki/Wikidata:How_to_use_data_on_Wikimedia_projects
7. https://www.wikidata.org/wiki/Wikidata:Wikidata_in_Wikimedia_projects#Wikidata_fed_infoboxes_on_Wikipedia
8. https://commons.wikimedia.org/wiki/Template:Creator
9. https://blog.wikimedia.org/2017/10/30/wikidata-fifth-birthday/
10. https://www.societybyte.swiss/2018/11/07/how-wikidata-is-solving-its-chicken-or-egg-problem-in-the-field-of-cultural-heritage/
11. https://pt.wikipedia.org/wiki/Wikip%C3%A9dia:Mbabel
12. https://meta.wikimedia.org/wiki/Wikidata/Notes/Article_generation
13. https://arxiv.org/pdf/1906.01965.pdf
14. https://www.sciencedirect.com/science/article/pii/S1071581917300642
15. https://ieeexplore.ieee.org/abstract/document/7427729
16. https://yukaichou.com/chou-musings/five-examples-of-gamified-crowdsourcing-to-learn-from/
17. https://tools.wmflabs.org/wikidata-game/
18. https://tools.wmflabs.org/wikidata-game/distributed/
19. https://www.researchgate.net/publication/335223330_The_Exploration_Game_on_Wikidata
20. https://www.wikidata.org/wiki/Q26919966
21. https://bitbucket.org/magnusmanske/wikidata-game/src/master/public_html/distributed/
22. http://magnusmanske.de/wordpress/?p=362