Problem outline and defining expected outcome
Version 2.0

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

The problem statement seeks to enhance Hindi or Telugu wikidata. Each wikidata entry requires a reference. Any webpage on the internet could be used as a potential reference for a wikidata entry. Now about 56% of the webpages are written in English while only 0.1% of the webpages are written in Hindi. This stats raises a serious concern which is lack of credible Hindi or Telugu sources which could provide data that could be used to enhance Hindi and Telugu Wikidata, which could be used to increase productivity of Wikipedia writers.

Now there are two ways to go ahead with this:

1. We scrape data from known/reliable places and find relevant information in it and ask users to verify it and if certain level of confidence is reached in the scraped data we use it to make wikidata edits.
2. We find out various properties of an item whose value is missing and we ask users to provide these values along with reliable sources. Once we reach a certain level of confidence in a proposed value we make wikidata edit.

Now a solution based on the first approach was earlier proposed but it had some glaring limitations (not scalable, messy game play, etc.) and thus in this version solution based on the second approach is presented. This solution builts on top of existing Wikidata games.

Problem Outline

1. **What is the problem?**
   I would define the problem as enhancing Hindi and Telugu Wikidata using gamification and then using that wikidata to enhance productivity of Wikipedia writers. Thus the first part of the solution would aim at enhancing Hindi and Telugu Wikidata and the second part will aim at providing users with this data so that they could use it to enhance wikidata. A common solution will achieve both the aims.

2. **Difficulties and its solution**
   As already mentioned in the starting that the amount of web pages in Hindi or Telugu are very less and thus accountable and factually correct source of information in these

languages are very less. Thus we aim to tackle this problem by seeking data/information from users and encouraging them to cite reliable sources.

3. **Wikidata domain that we are targeting**
   For this project wikidata entries on science and technology would be targeted. As of now following categories are targeted (non-exhaustive, more categories might be added later):
   a. Chemical Compound
   b. Chemical Elements
   c. Chemical Reactions
   d. Stars
   e. Theorems
   f. Diseases
   g. Softwares
   h. Biological Process
   i. Rivers
   j. Mountains
   k. Countries

4. **Target audience**
   The game is mostly meant for school students.

Expected Outcome

The outcome of the project would be a two subsystem game which would be able to:

1. Find out various properties of an item whose value is missing
2. Ask users to provide these values along with reliable sources
3. Reach a certain level of confidence on a proposed value.
4. Make Wikidata edits.

Solution Outline - Second Draft

Question Creator

Question creator finds out various properties of an item whose value is missing. Now if we look at all the categories that are targeted, all of them contain things that have almost similar properties. For example Chemical Elements contains Oxygen and Argon. If we look at Wikidata entries for both of them we will see that both of them have similar properties entries. For example both of them have Atomic mass, Atomic number, Discoverer, etc. Thus if some property is missing in Argon but is there in Oxygen then that property must be there in Argon too. For example - Oxygen has property “Oxidation state”, however this property is missing in Argon. Thus this system identifies such missing properties and then creates questions around them.

The game

1. The game first asks users to choose for the category they want to play and then from that category it identifies missing properties and formulates the question based on that missing property.

2. The user can choose to play in multiplayer scenario or single player scenario.

3. Now suppose the question based on missing property value is generated to be: “What is the oxidation state of Argon?” Now for this question we do not have options. Thus this is where multiplayer settings come in. In the multiplayer scenario, say 10 people contest together. They are given this question and asked for an answer and cite a source(not compulsory). If they are not sure about the answer they could choose to not answer the question. Suppose for the argon question, we get 3 answers. Six people(more than threshold) answered 0, One person answered 1 and Two people answered 2. Now clearly 0 is the answer. Now suppose someone has cited a source, then that person gets extra points (A lot of points and thus encouraging them to cite sources). Those who got the right answer get say 10 points. Those who cited sources get 100 bonus points (since the bonus is high, people are encouraged to cite sources). Those who answered wrongly get -10 points and those who choose not to answer get 0 points. This encourages people not to guess and answer only if they are sure. All this answering must be done in stipulated time, say 10 seconds.

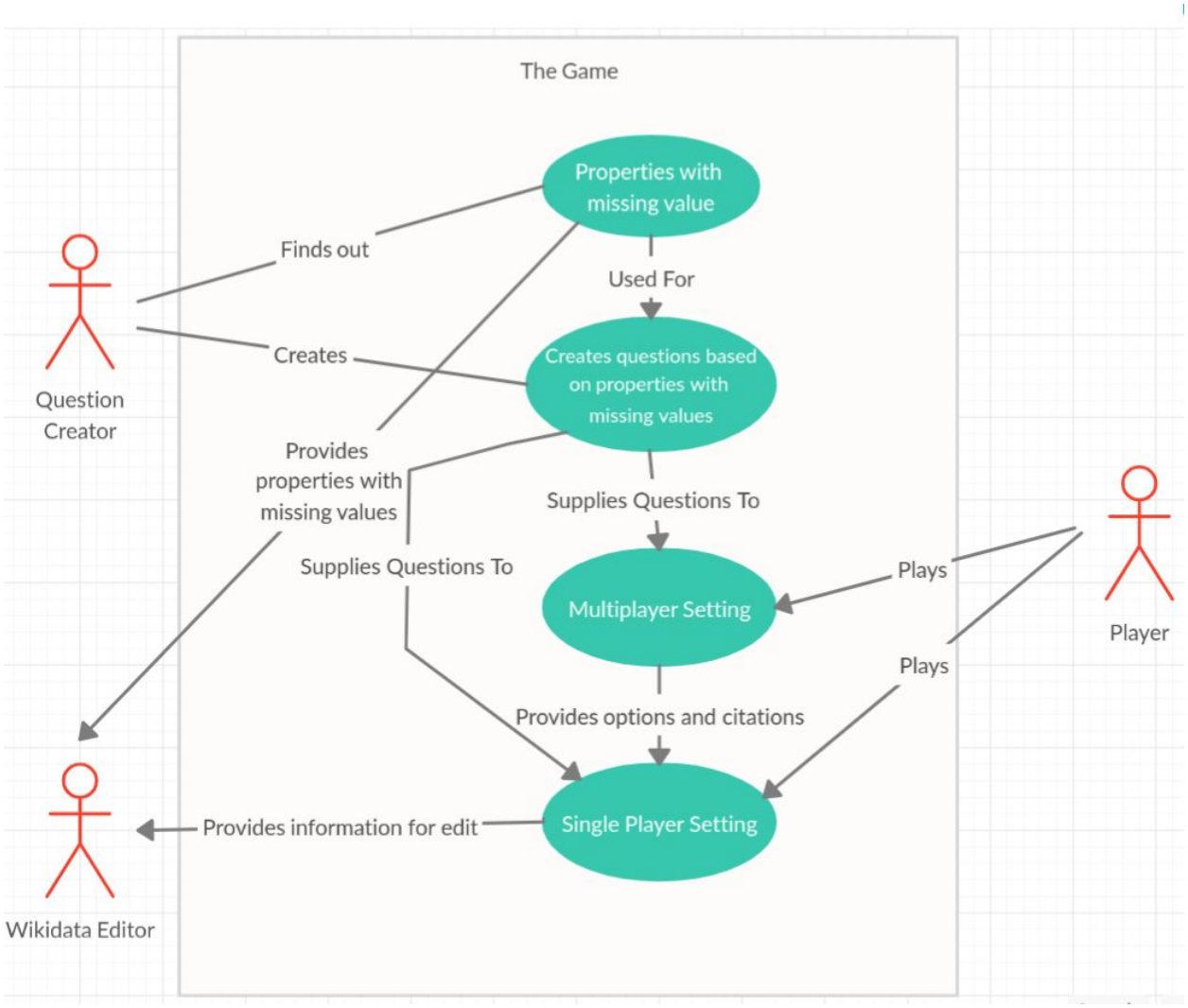
4. Now the majority is not always correct, so if someone has cited a source and is in the minority then his/her source is checked by a human and if the source seems to be reliable then his/her answer is taken to be the correct answer. Since human checking can take time points might be updated later and also we want to make it automated or semi-automated.

5. Now after multiplayer settings we have a question and options, we then ask users to answer this question and if in an answer certain level of confidence is reached we make corresponding updates in Wikidata. (This is mostly what the existing solution does)

Difference from existing solutions

Existing solution is like the single player scenario. However the game developer has to manually create options and thus scalability is low. However the proposed solution does not depend upon the developer to create/provide options, instead uses crowdsourcing for the same and actually encourages people to cite sources. The existing games were based on the belief that the majority are always correct, however this game lays more trust on sources than on what the majority believes.

Use Case Diagram


The Game
Properties with missing value
Finds out
Used For
Creates
Creates questions based on properties with missing values
Question Creator
Provides properties with missing values
Supplies Questions To
Supplies Questions To
Plays
Multiplayer Setting
Player
Plays
Provides options and citations
Provides information for edit
Single Player Setting
Wikidata Editor

❐ ❐ ❐