A Macro-Micro Biological Tour of Wikidata
Wikimania North America Preconference – Montreal - August 10, 2017

Slide No. 1: Overview

Thank you for joining me in this "Macro-Micro Biological Tour of Wikidata."

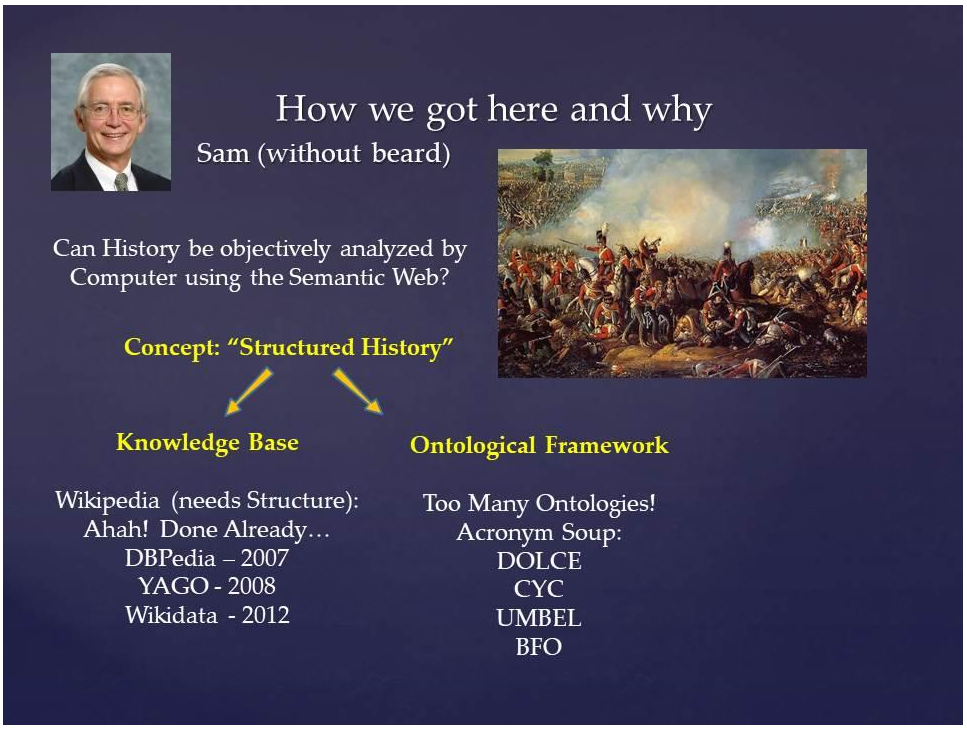
This tour will be a personal journey into the world of Wikidata to look at two extremes of living things – the Macro or Human scale, and the Micro or microbiological scale. I am most pleased to have as my traveling companion on this tour Dr. Yongqun He from the University of Michigan Medical Research Center. Yongqun goes by the name "Oliver" in the US, but currently he is in Beijing, keeping in touch by email and Skype.

I will first describe how this adventure was conceived, and then describe what we have found, provide the conclusions we have reached, and offer some recommendations for the future.

Slide No. 2: How We Got Here and Why (1)

My adventure began before I had this beard, which I have been growing in order to look like a pirate in our local community theatre production of the Pirates of Penzance in September.

In fact, my adventure began about two years ago when I made the following conjecture: There is an **objective reality underlying human history**, historical information is now in digital form, and current computer technology and emerging semantic web techniques should be able to analyze this information.



By doing so, it may be possible to accurately describe the causal factors. It may not be possible to show true cause and effect relationships, but it should at least be able to **disprove false narratives**. If so, could we potentially avoid some of the conflicts that have arisen from the false historical narratives of the past? From this perspective, I envisioned a project I am calling **the "Structured History" project**.

This project would need two things:

First it would need a **knowledge base** with access to historical information. I thought at the outset that my project would need to structure the data in Wikipedia.
However – I found that this has already been done, first by DBPedia, and more recently by Wikidata. There were also comprehensive alternatives, the most commonly found being YAGO developed by the Max Planck Institute.

Secondly, the project would need a complete, consistent and useable system of classification, or **Ontological Framework**, for organizing and analyzing the information. There are many ontologies to choose from – too many, it seems to me. How many ways should there be to organize our knowledge of the outside world?

Slide No. 3: How We Got Here and Why (2)

In searching for a good ontology, I thought it would be good to find one that had actually been used – an Applied Ontology that actually helped researchers do their work.

This led to the ontology framework that I believe has been the most widely used in multiple subject-oriented endeavors, which is an Upper Level Ontology called the **Basic Formal Ontology or BFO**. I also discovered that one of the experts in using BFO was only an hour's drive from my home, at the University of Michigan Medical Research Center. There, Dr. Oliver He has two laboratories, the first being what he calls his "wet lab" where he and his team actually work on bacteria in the lab, with a special focus on the disease Brucellosis and its related bacterium, *Brucella*.

His second activity is in bioinformatics, and he calls this his "dry lab." This lab has developed an ontology analysis service called "**OntoBee**," which is available for public use (http://www.ontobee.org). He also helps develop subject matter specific ontologies derived from BFO, and at the moment he is developing a cell line ontology under a grant from the US National Institutes of Health.

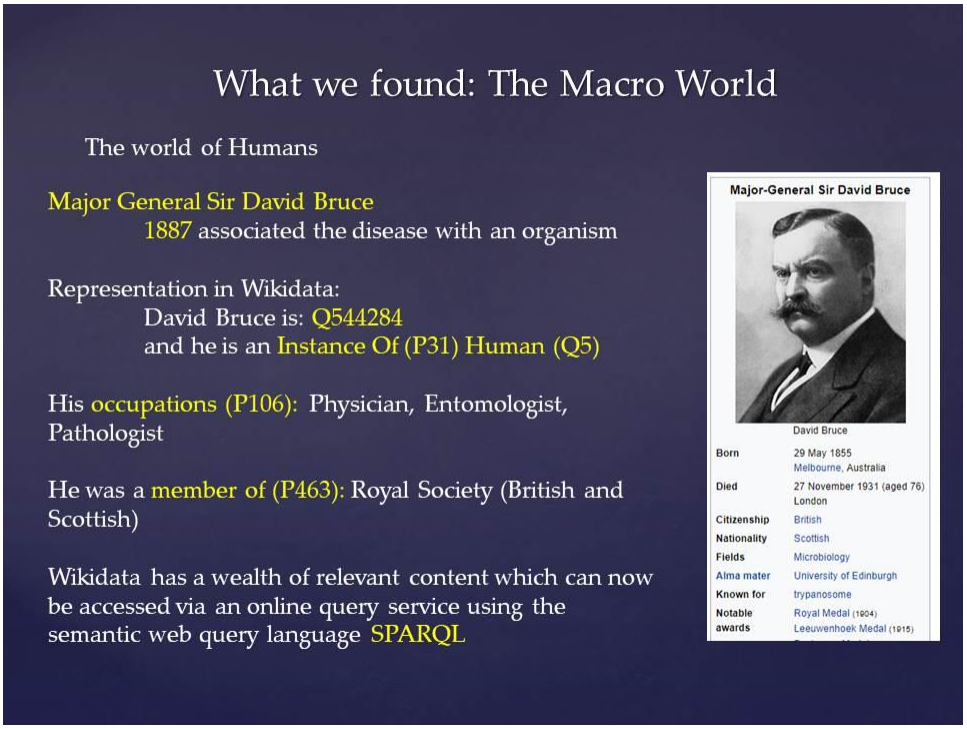
I was very pleased that Dr. He was interested in discussing our common interests at lunch over several months. I spoke highly of Wikidata as a growing knowledge base, and encouraged Dr. He and his colleagues to consider using Wikidata and possibly uploading their research results into this knowledge base.

From these discussions, we came up with the notion of what we called our "little project" – a **Macro-Micro Biological Tour of Wikidata** – a title that predated our knowledge that Wikimania was going to be in Montreal.

Slide No. 4. What we Found: Macro World

So my adventure began, looking first at the Macro World of humans, starting with the human responsible for identifying the Brucellosis disease, Major General Sir David Bruce, who first

associated the disease with an organism in 1887. There is a wealth of information in Wikidata about David Bruce and other pioneers in bacteriology.

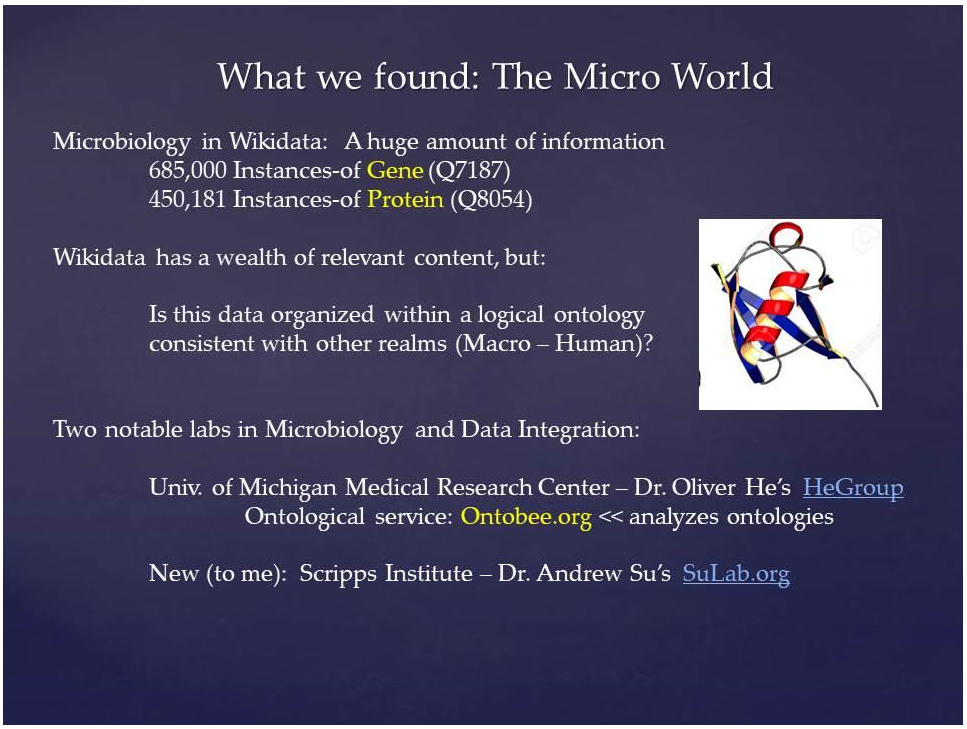


For those who have not yet used Wikidata extensively, I would like to show how Wikidata represents our Major General. Items about which statements are made are assigned a number prefixed with the letter Q, and properties about these items are enumerated with the prefix P. So Major General David Bruce is Q544284, and he is an instance-of (P31) Human (Q5) His occupations and professional memberships are as shown, with numbered properties.

Wikidata can now be accessed online via a powerful query language, SPARQL, the details of which are beyond the scope of this talk. But with SPARQL one may find all the people involved in bacteriology and sciences leading up to bacteriology.

Slide No. 5. What we Found: Micro World

If we turn our attention to the Micro world, there is a massive amount of information to be mined here, and my impression is that it has grown since I first looked about a year ago. There are now 685,000 things in Wikidata that are Instances-of a Gene, and 450,000 things that are Instances-of Protein.



However, while there is plenty of Micro data available, how well is this data characterized in Wikidata? This is a topic we will cover shortly.

I was really pleased to find out about Dr. He and his activity at the University of Michigan, with the links to his HeGroup (http://www.hegroup.org) indicated. However, a recent "find" for me was an activity led by Dr. Andrew Su at the Scripps Institute in California (http://sulab.org) [his work is well known to Wikimedia insiders, but news to me].

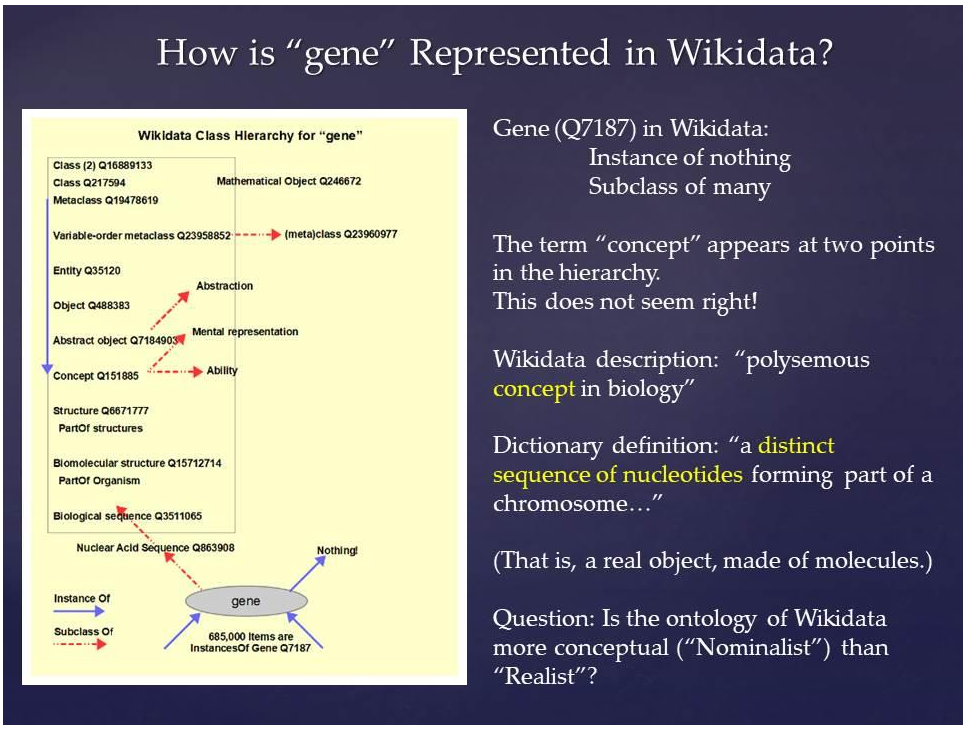
Slide No. 6. What we Found: Micro World

The reason I was excited to find out about SuLab stems from a comment you may recall from slide 3 – where I encouraged Dr. He to use Wikidata. Well, it turns out that SuLab has done just that, and he has uploaded his results to Wikidata.

As stated in the article shown (http://sulab.org/2017/07/the-gene-wiki-project-looking-to-the-future-v-2017), he states that "Our team was the first to perform systematic loading of biomedical data in Wikidata." Dr. Su was nice enough to have an email exchange and to approve the information on this slide. I mentioned the power of SPARQL queries against Wikidata, and I would highlight the link to SuLab.org that contains a wealth of sophisticated query examples (https://www.wikidata.org/wiki/User:ProteinBoxBot/SPARQL_Examples).

Slide No. 7. How is a "gene" Represented in Wikidata

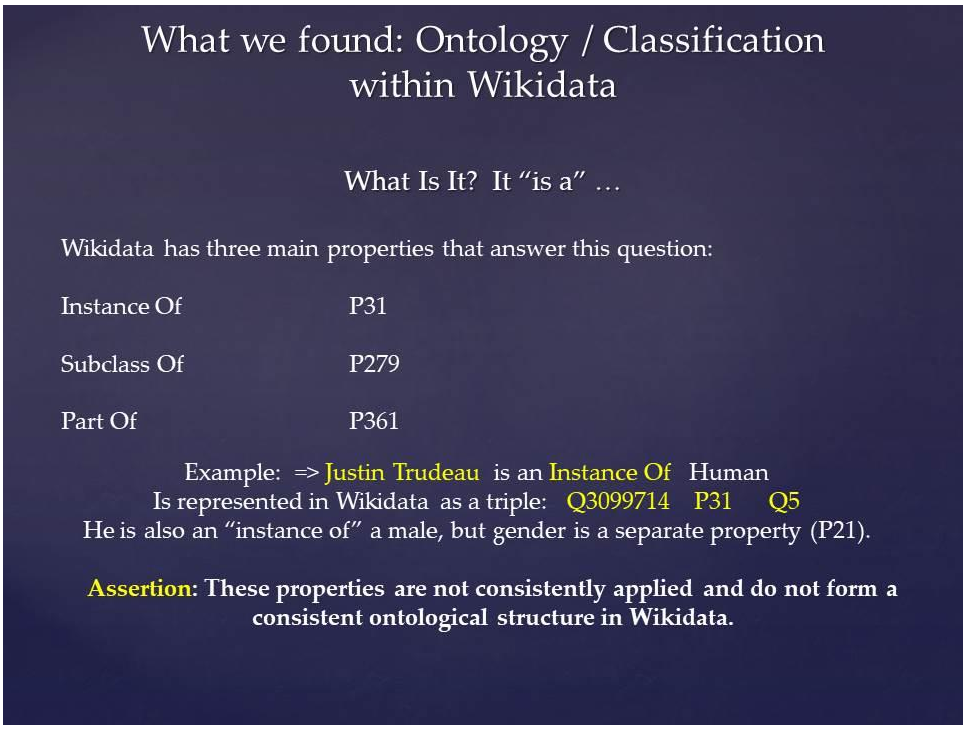
In looking into the Macro and Micro worlds, I wanted to see how various entities were classified within Wikidata. This chart shows the hierarchy of classes to which the term "gene" belongs. The two relations of interest are "Instance Of" and "Subclass Of." Surprisingly, gene is not an instance of anything, but it is a subclass of a sequence of more general terms. However, while the chart is necessarily small print to capture all the terms and may not be readable, I wanted to reflect that it goes up through the term "concept," but continues up to a higher level where "concept" appears again. It is ultimately a subclass of "variable-order metaclass." Whatever that is. It is a lot of stuff, involving "mental representation" and "abstraction."



Another aspect of the representation of "gene" within Wikidata is that it is described as a **concept**: "A polysemous concept in biology," indicating the term has several meanings. The online dictionary, however, is much more tangible: "a distinct sequence of nucletides forming part of a chromosome…" The definition is "real" – made of molecules, whereas Wikidata emphasizes the conceptual nature of the term. I am not an ontological specialist, but would it be true, in this case, that Wikidata is showing a "Nominalist" view of nature, rather than a "Realist" view?

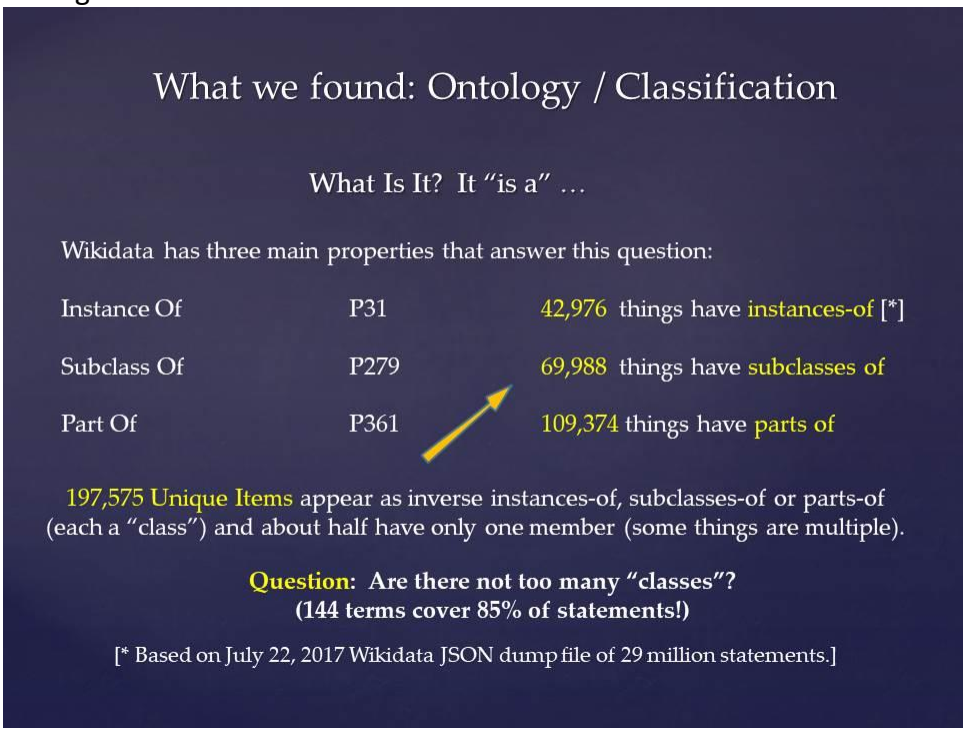
Slide No. 8: What we found: Ontology / Classification within Wikidata (1)

The classification terminology within Wikidata hinges on three significant properties, Instance-of, Subclass-of, and Part-of, designated P31, P279 and P361 respectively. For instance, the assertion that "Justin Trudeau is an instance of Human" would be a triplet statement, connecting the Q number for Trudeau to the Entity Human by the property P31, Instance-of. I would assert that these properties are not applied consistently within Wikidata.



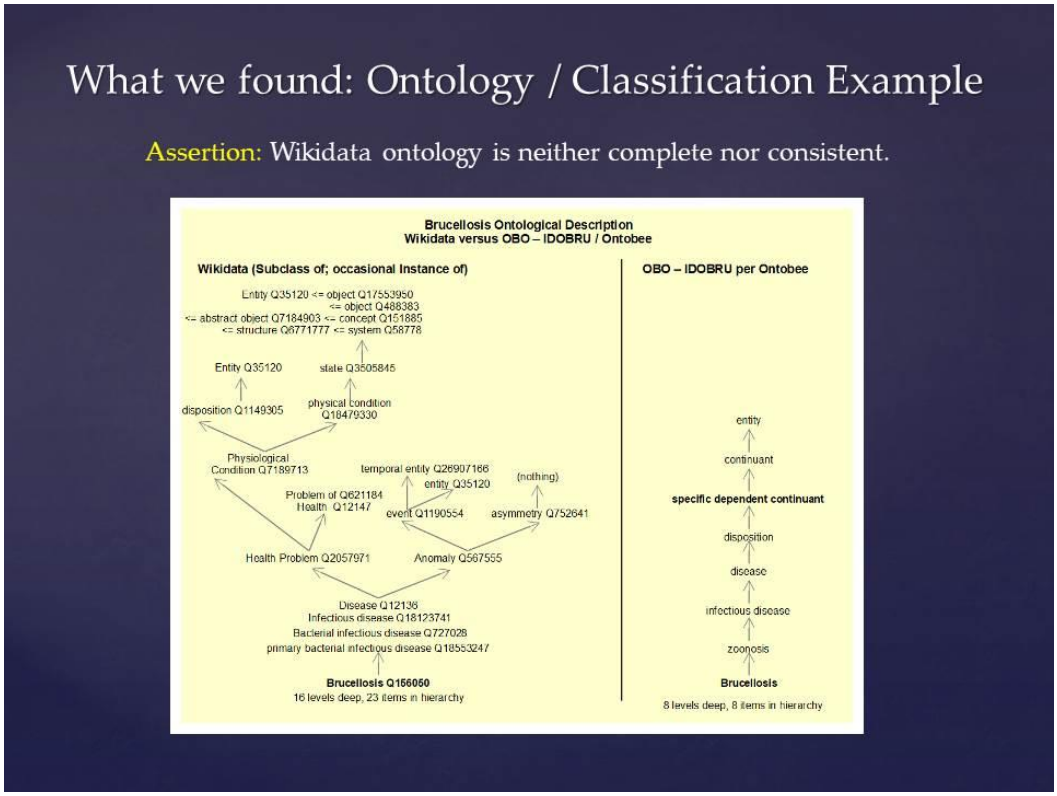
Slide No. 9: What we found: Ontology / Classification within Wikidata (1)

Additionally, the number of terms used to describe things is extremely high – over 197,000. Of the 29 million items with statements in Wikidata, 85 % of them are described by only 144 terms, meaning there almost 200,000 terms to describe 15% of Wikidata contents. This seems like too many, especially considering that almost half of statement items are single items within a class – they are essentially instances of themselves. Culling and curation of important classifying terms would seem like a good idea for serious research.



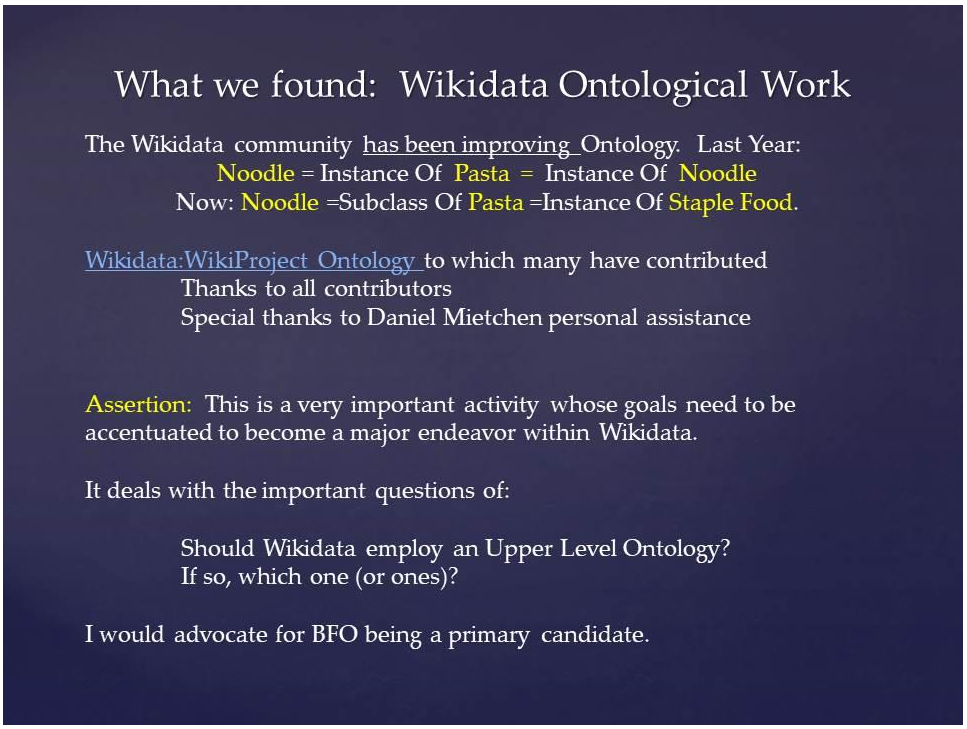
Slide No. 10: What we found: Ontology / Classification Example

Another example shows the classification of Brucellosis in Wikidata compared to the handling with the **Open Biological Ontology (OBO)** that is patterned after BFO. On the left you can see the pathway up through multiple entities in Wikidata, contrasted with the streamlined, non-branching classification within OBO.



Slide No. 11: What we found: Wikidata Ontological Work

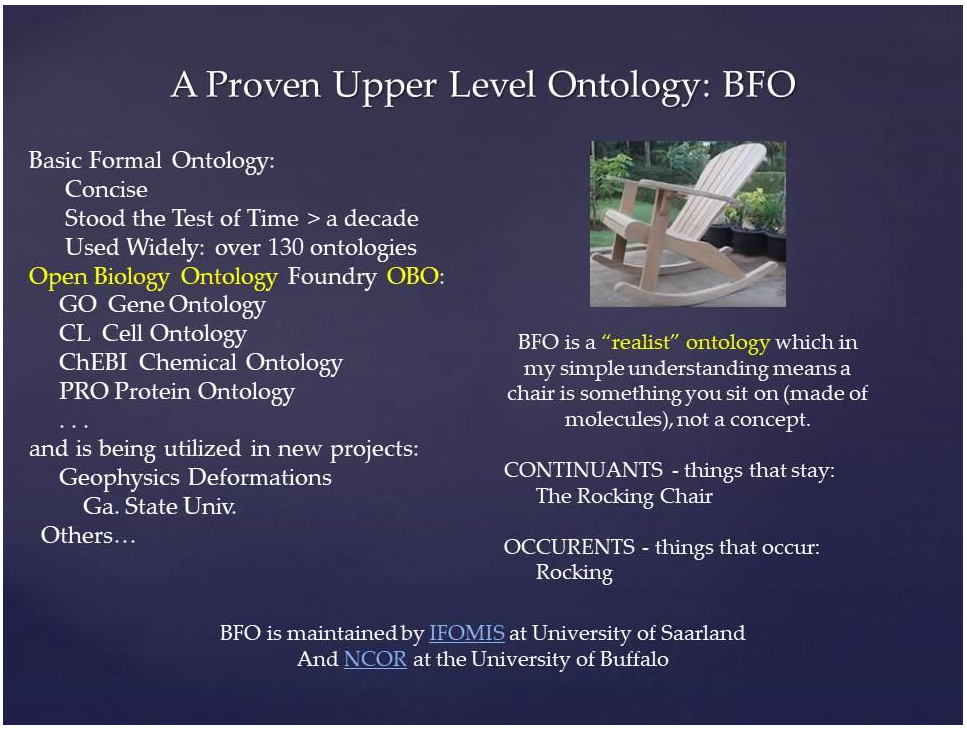
Over the last year, the Wikidata community has been improving its ontological treatment of the content. A year ago a Noodle was an instance of a Pasta, which in turn was an instance of Noodle! This is now more properly treated with noodle being a subclass of pasta (Oops: Its definition was just changed again on August 7 as I'm writing this!) Wikidata content is definitely dynamic and realtime!



There is an existing project within Wikidata on Ontology to which many people have contributed (https://www.wikidata.org/wiki/Wikidata:WikiProject_Ontology). I would like to take this opportunity to thank the contributors, and thank Daniel Mietchen for his personal correspondence to give insight into the Wikidata community. I will assert that this project is imporant and warrants additional emphasis and support. Its mission includes many of the important questions regarding the merit of an Upper Level Ontology, and a comparative assessment of the major candidate frameworks.

Slide No. 12: A Proven Upper Level Ontology: BFO

I would like to accentuate the reasons I am advocating BFO as a good candidate among ontologies. It is concise, it has stood the test of time for over a decade, and it is widely used, with over 130 derived subject ontologies, a few of which are indicated. In the last week I have learned of a geophysical deformation ontology being developed at Georgia State University, and a geological-historical ontology from the University of Lublin in Poland that will be using BFO.

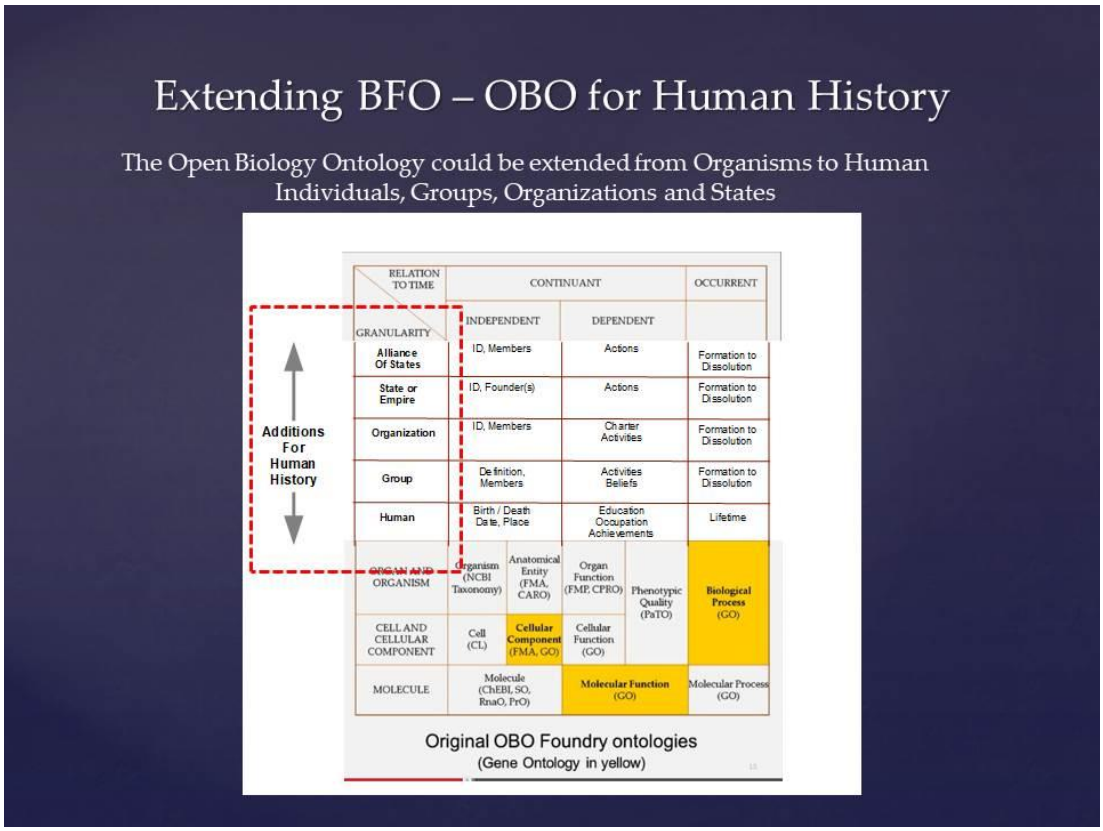


BFO it referred to as a "realist" ontology, rather than the alternative – meaning it deals with things not concepts. Things are categorized as Continuants that do not change with time, and Occurants, things that do change. For instance, **the rocking chair** is a continuant but **rocking** is an occurant.

BFO is actively maintained and promoted by two groups, IFOMIS (http://ifomis.uni-saarland.de/bfo) in Germany and the US **National Center for Ontological Research (NCOR)** (http://ncorwiki.buffalo.edu/index.php/Basic_Formal_Ontology_2.0).

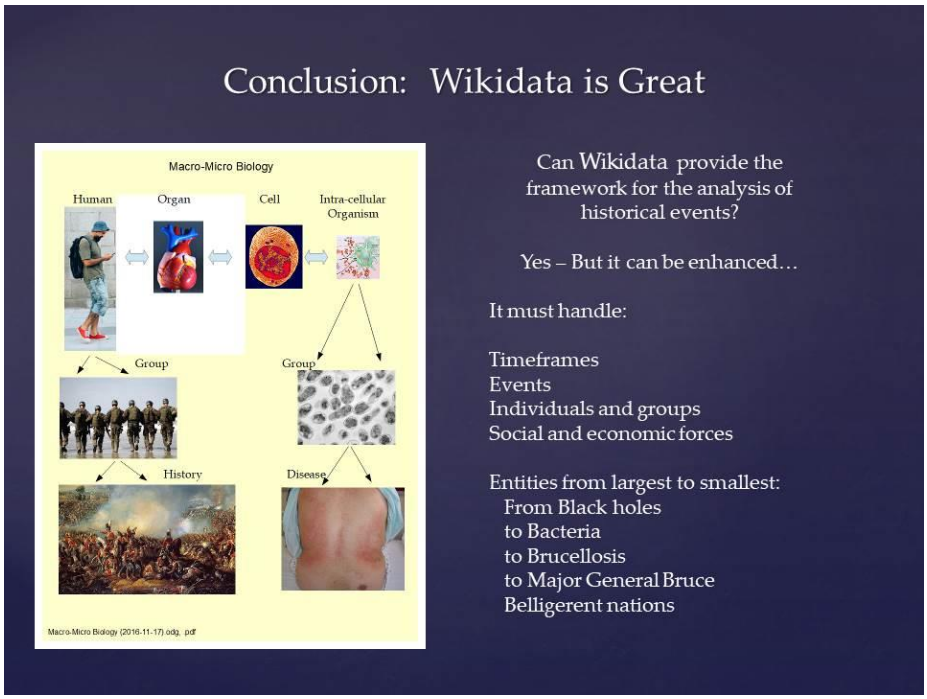
Slide No. 13: Extending BFO – OBO for Human History

The basic framework of BFO as used in the Open Biology Ontology could be readily extended to Humans, covering individuals, groups of individuals, organizations and states. This chart shows how these new entities may appear in an expanded OBO framework.



Slide No. 14. Conclusion: Wikidata is Great

My Macro-Micro Biological Tour of Wikidata leads me to conclude that Wikidata is GREAT! However, I think it can be enhanced and needs to be enhanced to serve as a research-ready resource. For Macro-Micro Biology synthesis, it needs to handle: Timeframes, events, individuals and groups as well as social and economic forces. And it need to do this in a consistent manner across a vast range of entity sizes, from Black holes to Bacteria, from Brucellosis to Major General Bruce, to analyzing Beligerent nations.



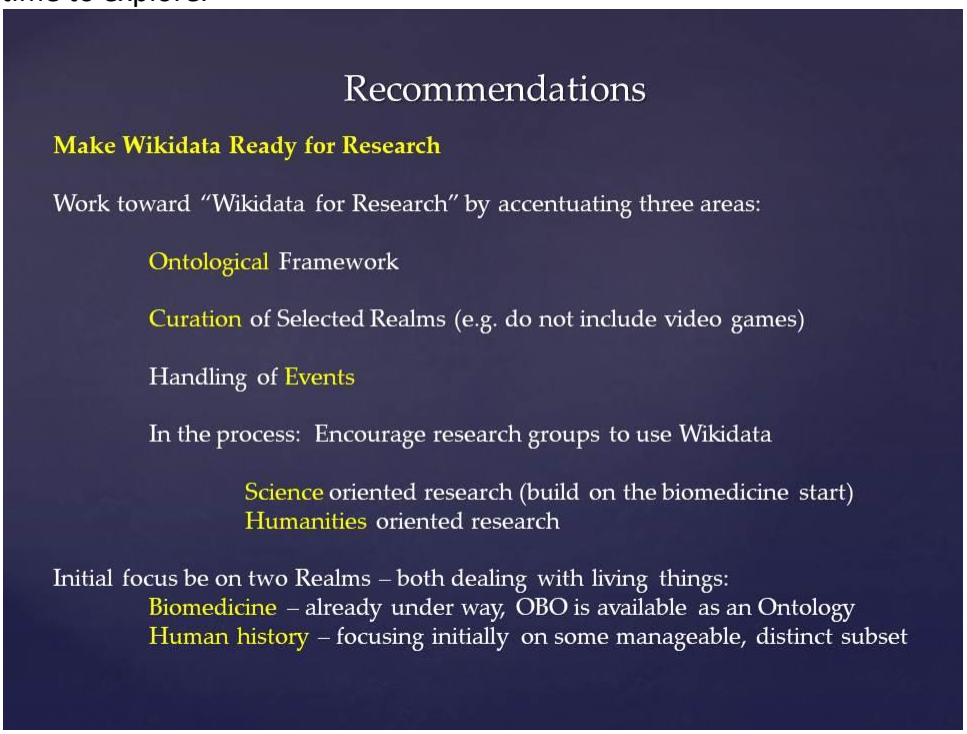
Slide No. 15: Macro-Micro History

The prospect that I would like to hold out is the ability to find, correlate and analyze historical information of diverse types, showing what happened and possibly why. A hypothetical example is shown here, but be able to the track the migration of a disease and and associate this phenomenon with the movements of the 16th Roman Legion. That would be digital history on steroids!

Slide No. 16: Recommendations

My recommendation is to intensify the enhancement of Wikidata to make it **ready for research**. Three areas I would accentuate are:

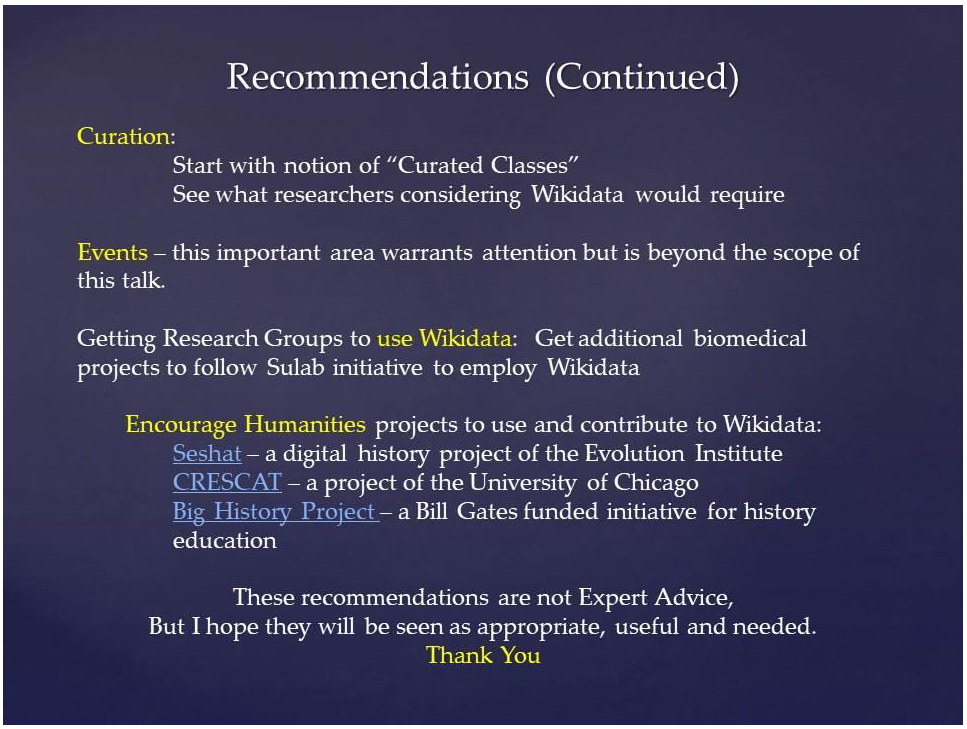
1. An enhanced **ontological framework**,
2. The **curation** of selected classes as a means of quality control, and
3. **Improved handling of events**, an important topic for history and other fields that we do not have time to explore.



In the process, it would be good to encourage research groups to follow the lead of SuLab by uploading their results to Wikidata. These groups could include not only scientific groups but also research projects in the humanities. My suggestion is to start with two realms, biomedicine and human history.

Slide No. 17: Recommendations (Continued)

In a parallel activity, the issue of curation could be explored, beginning with a dialogue with researchers such as Dr. Su who plan to use Wikidata, to assess their needs. A suggestion is to accomplish this by **designating some classes as "Curated Classes"** for which the content is monitored to ascertain sufficiency and validity. Dr. Su discusses these issues in the link on the SuLab slide.



An important area for history and other dynamic processes is the **handling of events**. This is an area outside the scope of this talk, but an impression is that this area needs attention. One key factor is, what constitutes a significant event, since the granularity of events is unlimited. That is, each historical person could potentially have a event of some note each hour, but how many need to be stored? [A notable paper that came out the week of Wikimania is **The Rich Event Ontology**, which discusses the issues underlying event ontologies and a proposed system for handling. (http://aclweb.org/anthology/W17-2712)]

Lastly, it would be helpful to encourage both scientific and humanities projects to consider using Wikidata as both their source and their repository. For instance, at least three digital history projects are under development at this time that may be candidates:

Seshat (https://evolution-institute.org/project/seshat),
Crescat (https://oi.uchicago.edu/article/ochre-highlighted-rcc-article), and
The Big History Project (https://www.bighistoryproject.com/home).

These recommendations are from a user perspective and not intended to be expert advice. But I hope they wil be seen as reasonable, appropriate, useful and needed. Thank you.