UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département d'astronomie

ISDC

Knowledge Graphs

Ecogia Science Meeting

Volodymyr Savchenko

November 8, 2020

Why is this interesting?

This is not a (natural) science topic, but it's useful for scientists, since **scientists deal in knowledge**

Not new or revolutionary technology, and does not generally replace other valuable data management technologies.

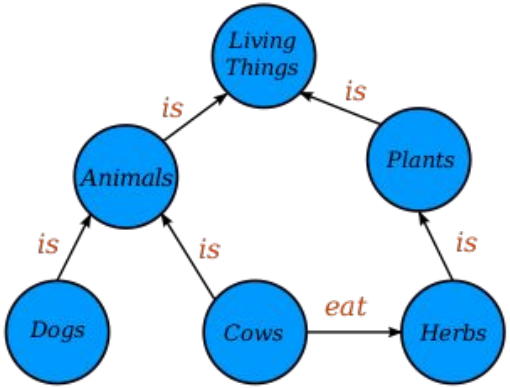
It helps to think clearly of **common terminology** and precision of expression, and exposes **challenges** of this process. **Necessary component of open data world!**

It's interesting as has to do with **how the web was built**, organized, and how it's becoming

While the topic is old (from 197x) and is not widely known for some good reasons, but **recent iteration (since ~2017) gained a lot of traction**

I am not an expert, just trying to share my domain-linked experience!

Knowledge Graph databases



Can be seen as collection of **triples**, as:

:Herbs :are :Plants

Initially loosely shaped structure, then constrained with an **ontology**, e.g. (simplified)

:LivingThings :mustHaveProperty :eat

Can be enhanced with **inference rules** (see full necessary graph on the left):

:Cows :eat :Plants => :Cows :eat :LivingThings

This information can be stored in a relational database, but a graph just makes more sense

WWW

WWW (not the internet, dns, css, javascript) was created in 1989 by *Tim Berners Lee* at *CERN* as several components, most significantly:

HTTP: protocol for fetching data

HTML: “standard” representation language

This also implied **URLs** as globally unique resource locators (or identifiers)

Interesting to note that it is defined as set of known “**recommendations**”, **RFC** (request for comments).

Browser developers surprisingly mostly follow them - proof that good open common language may be even commercially favorable to follow

WWW Consortium (**w3c**) continues to issue RFC.

WWW towards Semantic Web (2001)

As WWW grew, it become apparent that linking together global resources referencing poorly structure blocks of text, images, diverse tables does not allow to effectively deduce and find what the web contains.

W3c suggested new format, **RDF**, to represent web data as propositions, triples, reflecting relations between **URI**:

<http://www.wikidata.org/wiki/Cow> rdfs:type <http://www.wikidata.org/wiki/Animal>

RDF documents are themselves retrievable by **HTTP**

This allows to make the web "**Semantic**": relying on structured information exchange, not data blocks. **Make web content expreess statements about other web content.**

Since **URI** are global, web can be queried as structured global database of facts and data: "Linked Data" paradigm.

RDF is a not itself a language, and can be represented in multiple different languages (turtle, json-ld, etc)

Google Knowledge Graph

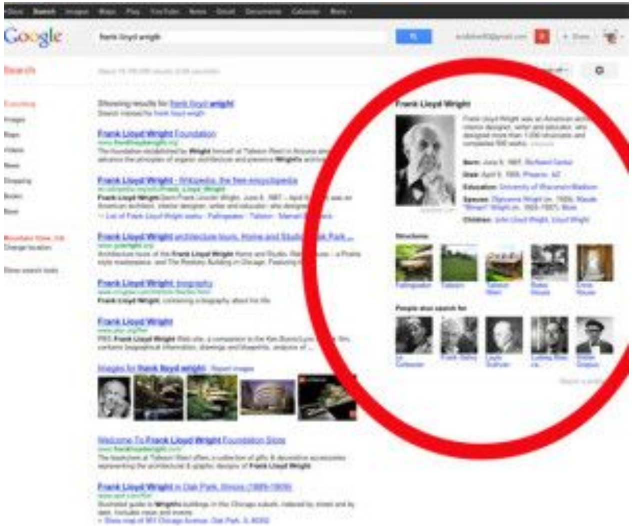
Linked Data and Semantic Web did not take off, since web developers did not like it, and market pressure did not favor it.

Another approach proven to be successful: building upon existing **Web of text** and creating **private structured Web representation: e.g. Google Knowledge Graph**

By May 2020, this had grown to **500 billion facts** on 5 billion entities, and consumption of **facts from the Graph exceeded clicks on regular results**.

Since ~2017, new forces emerged in business and academia, requiring open sharing of structured data. Also **technologies** allowing to do this by relatively small developments **became available**.

Knowledge Graph Facts here:



Growth of open Knowledge Graphs

https://www.wikidata.org

Growth started as **bottom-up initiative**. I was surprised to discover how much it got adopted without any formal requirement.

Businesses started to realize that open **global fact database** is useful to gain competitive insights.

Academics started to appreciate that **FAIR** can not be implemented without sufficiently powerful formal **common language**

Publishers, and NLP (un-publishers), naturally quite took up this topic, but they do not sufficiently understand the domains, their fact stores are limited

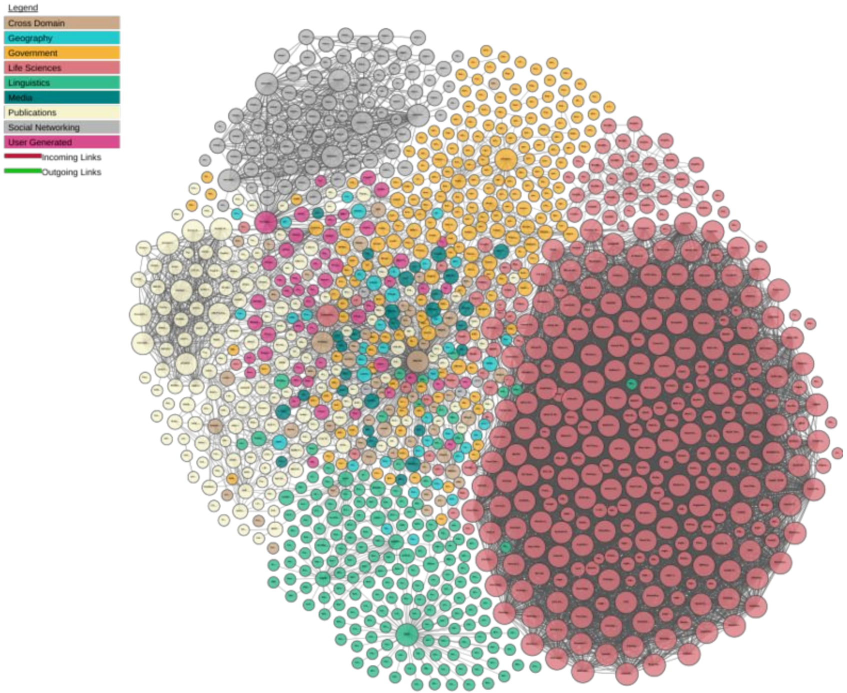
https://schema.org

https://data.crossref.org/10.1051%2F0004-6361%2F202037850

https://graph.openaire.eu/

https://lindas.admin.ch/

https://www.w3.org/wiki/SparqlEndpoints



International Virtual Observatory Alliance (**IVOA**)

https://www.ivoa.net/rdf/

Astronomy deals with common entities. As it often happens, it was pioneering in adopting **KG**s (but did not get very far). **IVOA** largely followed the approach of **w3c**

CDS/Simbad took care of creating with reference of object names and **object types**, often with **RDF URIs.**

Sesame is a great collection of **table data**, but with only partial semantic annotations: Researches often do not care to speak in common terms CDS did not manage all.

IVOA WGs regularly holds extensive discussions on improving **RDF** data model for astronomical entities.

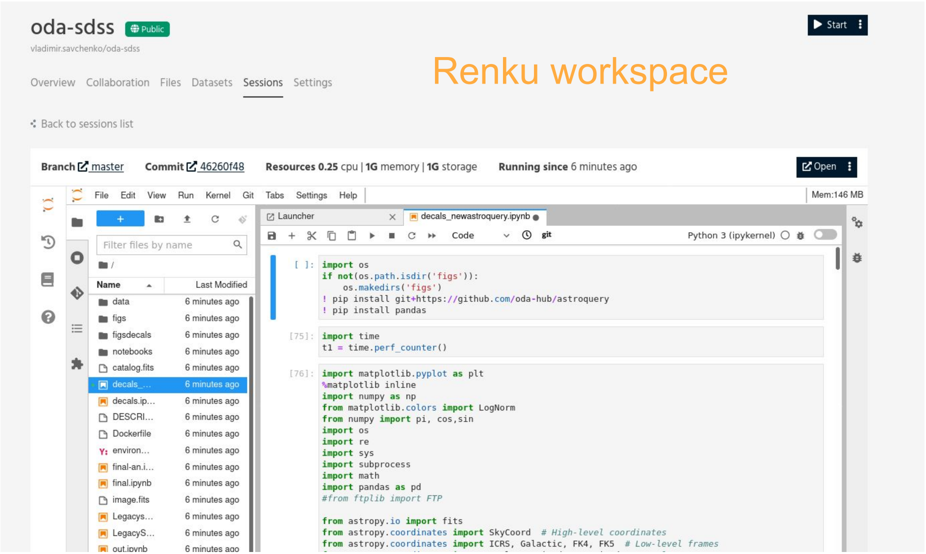
Natural sciences do not have large enough global KG, unlike humanities.

EPFL, SDSC, Renku

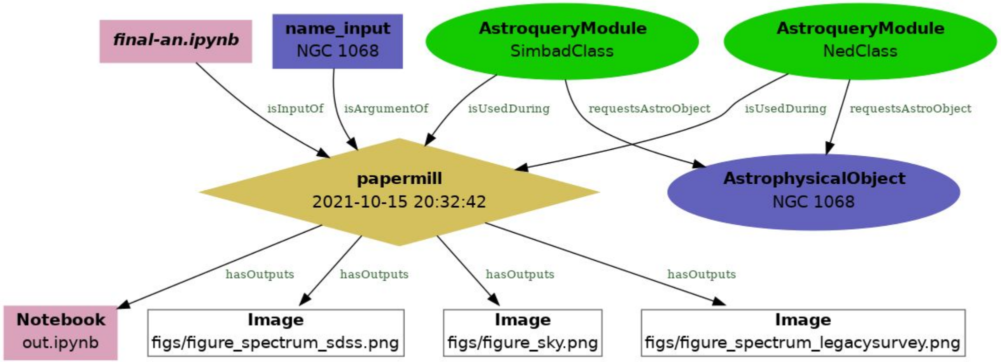
Renku offers space for collaborative research (like **jupyterhub, sciserver,** etc), but also helps users to implicitly build and leverage **knowledge graph**

EPFL-UNIGE project with **A. Neronov** and **G.Barni** improves features of renku relevant for building of astronomical **KG**.

This can eliminate the effort needed to create an **RDF description** of a scientific workflow and embed it in the **Linked Data World**



Graph description of the workflow harvested from Renku



Multi-messenger

We parse information from transient alerts into internal Knowledge Graph, and

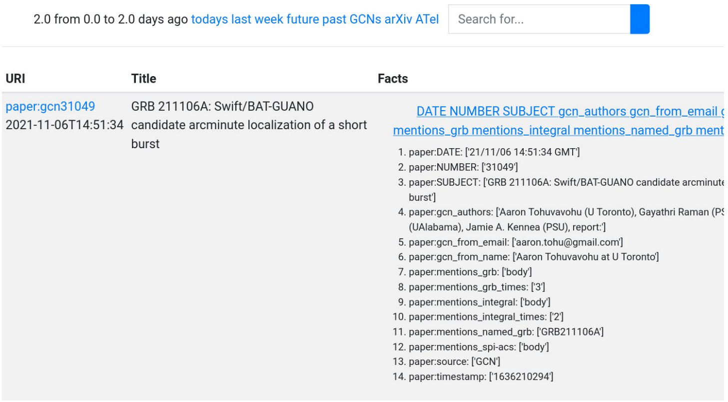
It’s convenient that not only the data but also the structure is not rigid and evolves easily. Inference rules allow to naturally express fact transformations.

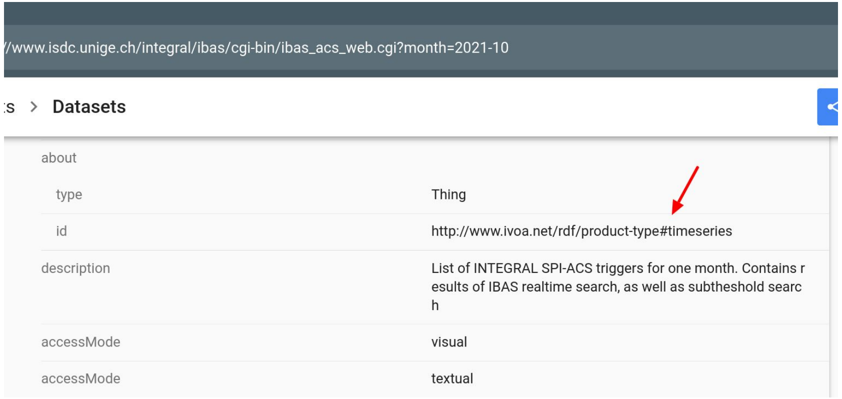
:GRB170817A: :is :shortGRB

=>

:GRB170817A: :isVeryRelevantFor :GravitationalWaveSearch

We publish some results in RDF or embedded in HTML, and want to do it more.

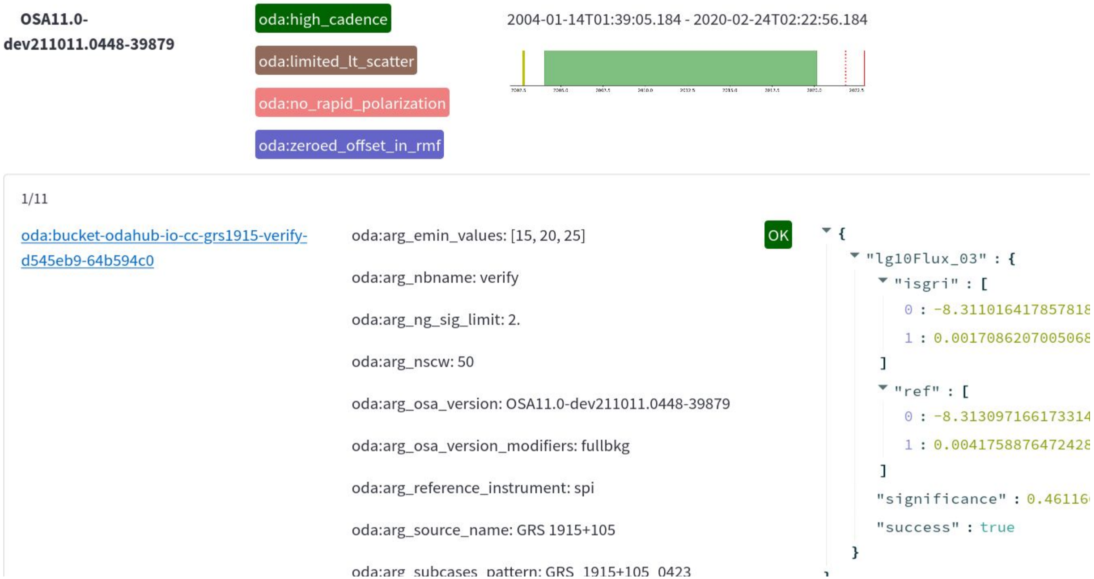




INTEGRAL cross-calibration

Cross-calibration between INTEGRAL ISGRI and various **“reference” information** naturally relies on global linked facts and concepts.

Using, among other, IVOA ontologies. So far, it was essential for natural organization of internal activities, integrated them.



When and how this can be useful?

Knowledge Graphs **essentially consist of citations** to external sources: citing is a good practice and allows to for easier tracing **what we consume**. These citations are not mere references, they (so good and bad citations can be distinguished)

Publishing results by embedding **common concepts** and **structured facts** enhances **clarity** and may benefit re-use, even if it does not replace human-readable text.

May help to **trace impact of publishing**, helps to link justification of decision making to scientific products, reduce opaqueness and misrepresentation of scientific studies

Enables to **model inference**, a form of **AI** (see recent PhD defense in UNIGE), allowing to automatically **validate and deduce structured propositions**

Empowers **machine learning on the graphs**, rapidly growing topic **enhancing inference**

Unlike humanities or bioinformatics, in **astronomy**, despites efforts of IVOA, the amount of information is not sufficient to leverage existing **KG**s. It does seem to be changing, under **pressure from FAIR** science.

Even for **local project knowledge base** it appears to be a great reasonably solution, the one which approaches global tipping point.