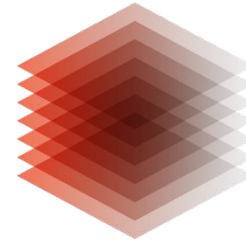


LEIBNIZ INFORMATION CENTRE
FOR SCIENCE AND TECHNOLOGY
UNIVERSITY LIBRARY



TIB

Open Research Knowledge Graph - A Lighthouse in the Publication Flood -

Anna-Lena Lorenz

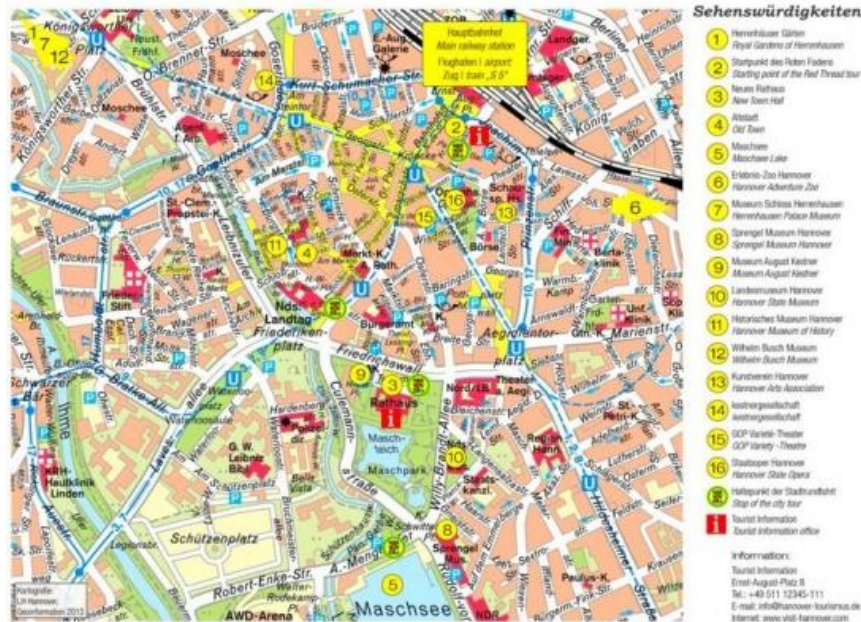
Coffee Lecture



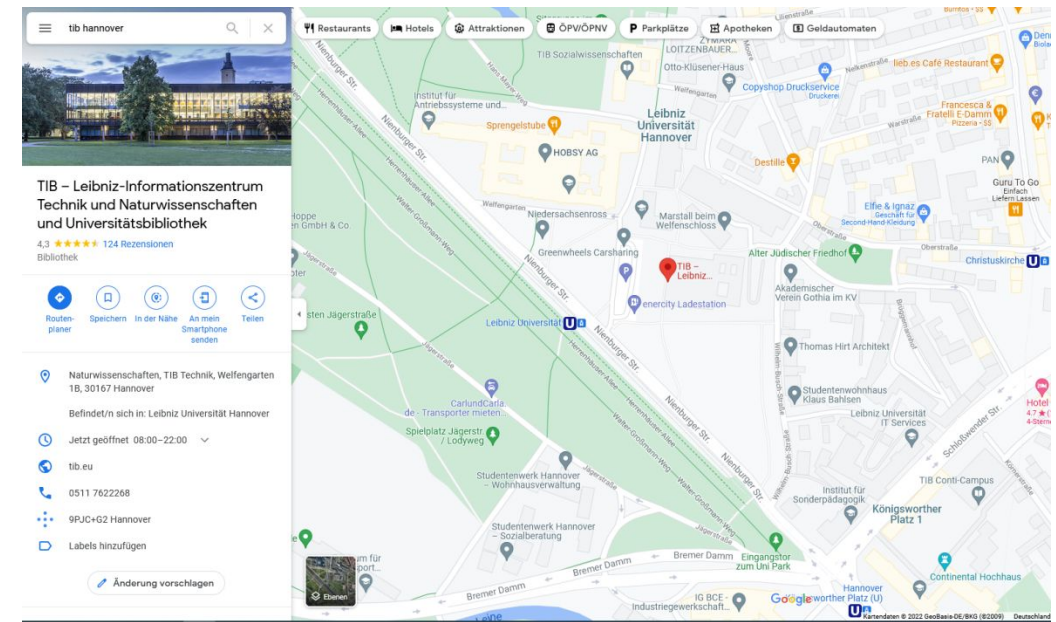
Creative Commons Attribution 3.0 Germany
<https://creativecommons.org/licenses/by/3.0/de/deed.en>

Digitalization in everyday life

Navigation to TIB Hannover
50 years ago



Now

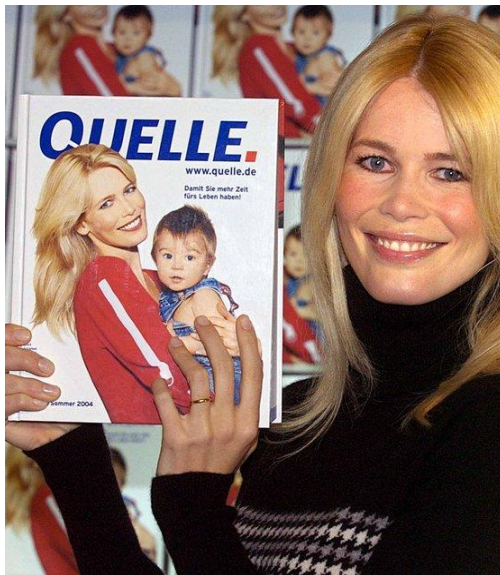


+ New Features:

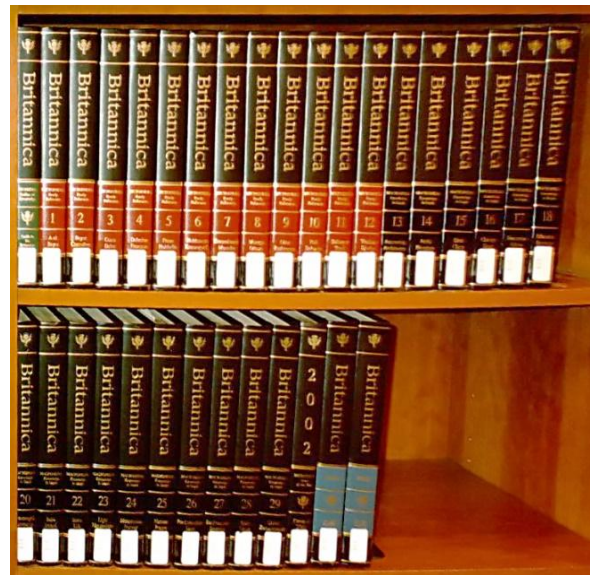
- Zoom in
- Traffic jam warning
- Opening hours
- Interesting places around

Similar in other domains....

Who still remembers?



Mail order catalogs



Encyclopedias

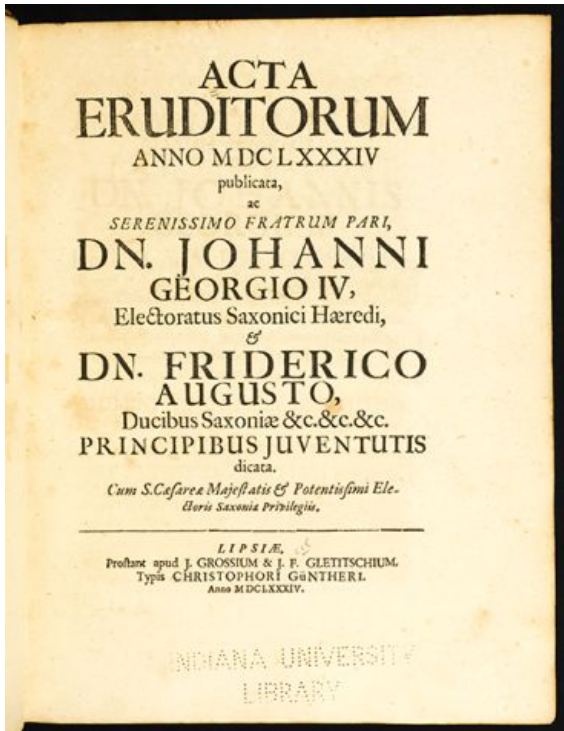


Phone books

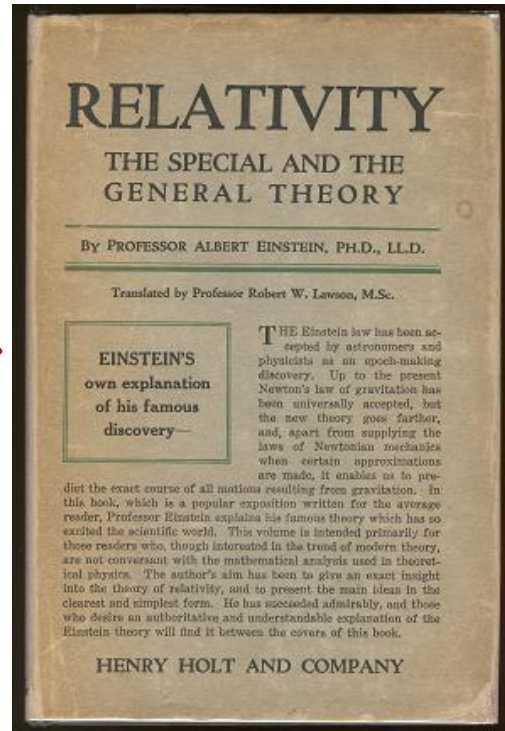
Whole industries got disrupted and our lives were significantly changed

What about Science?

Over 300 years ago



100 years ago



20 years ago

Information Retrieval

A Relational Model of Data for Large Shared Data Banks

E. F. Codd
IBM Research Laboratory, San Jose, California

Future users of large data banks must be protected from having to know how the data is organized in the machine (the internal representation). A prompting service which supplies such information is not a satisfactory solution. Activities of users at terminals and most application programs should remain unaffected when the internal representation of data is changed and even when some aspects of the external representation are changed. Changes in data representation will often be needed as a result of changes in query, update, and report traffic and natural growth in the stores of information. Existing noninferential, formatted data systems provide users with tree-structured files or slightly more general network models of the data. In Section 1, inadequacies of these models are discussed. A model based on *n*-ary relations, a normal form for data base relations, and the concept of a universal data sublanguage are introduced. In Section 2, certain operations on relations (other than logical inference) are discussed and applied to the problems of redundancy and consistency in the user's model.

KEY WORDS AND PHRASES: data banks, data base, data structure, data organization, hierarchies of data, networks of data, relations, derivability, redundancy, consistency, composition, join, retrieval language, predicate calculus, security, data integrity.

1. Relational Model and Normal Form

1.1. ISYMNOSUCTUS

This paper is concerned with the application of elementary relation theory to systems which provide shared access to large banks of formatted data. Except for a paper by Childs [1], the principal application of relations to data systems has been to deductive question-answering systems. Leavin and Maron [2] provide numerous references to work in this area.

In contrast, the problems treated here are those of data *independence*—the independence of application programs and terminal activities from growth in data types and changes in data representation—and certain kinds of data inconsistency which are expected to become troublesome even in nondeductive systems.

Volume 13 / Number 6 / June, 1970

P. BAXENDALE, Editor

The relational view (or model) of data described in Section 1 appears to be superior in several respects to the graph or network model [3, 4] presently in vogue for non-inferential systems. It provides a means of describing data with its natural structure only—that is, without superimposing any additional structure for machine representation purposes. Accordingly, it provides a basis for a high level data language which will yield maximal independence between programs on the one hand and machine representation and organization of data on the other.

A further advantage of the relational view is that it forms a sound basis for treating derivability, redundancy, and consistency of relations—these are discussed in Section 2. The network model, on the other hand, has spawned a number of confusions, not the least of which is mistaking the derivation of connections for the derivation of relations (see remarks in Section 3 on the "connection trap"). Finally, the relational view permits a clearer evaluation of the scope and logical limitations of present formatted data systems, and also the relative merits (from a logical standpoint) of competing representations of data within a single system. Examples of this clearer perspective are cited in various parts of this paper. Implementations of systems to support the relational model are not discussed.

1.2. DATA DEPENDENCIES IN FINANCIAL SYSTEMS
The provision of data description tables in recently developed information systems represents a major advance toward the goal of data independence [5, 6, 7]. Such tables facilitate changing certain characteristics of the data representation stored in a data bank. However, the variety of data representation characteristics which can be changed without *logically impairing some application programs* is still quite limited. Further, the model of data with which users interact is still cluttered with representational properties, particularly in regard to the representation of collections of data (as opposed to individual items). Three of the principal kinds of data dependencies which still need to be removed are: ordering dependence, indexing dependence, and access path dependence. In some systems these dependencies are not clearly separable from one another.

1.2.1. Ordering Dependence. Elements of data in a data bank may be stored in a variety of ways, some involving no concern for ordering, some permitting each element to participate in one ordering only, others permitting each element to participate in several orderings. Let us consider those existing systems which either require or permit data elements to be stored in at least one total ordering which is closely associated with the hierarchical listing or ordering of addresses. For example, the records of a file concerning parts might be stored in ascending order by part serial number. Such systems normally permit application programs to assume that the order of presentation of records from a data file is identical to (or is a subordering of) the

Communications of the ACM 377

Today

BIBLIOTHEK - Forschung und Praxis 2020, 44(3): 316-329

DE GRUYTER

Textmining

Sören Auer*, Allard Oelen, Muhammad Haris, Markus Stocker, Jennifer D'Souza, Kheir Eddine Farfar, Lars Vogt, Manuel Prinz, Vitalis Wiens and Mohamad Yaser Jaradeh

Improving Access to Scientific Literature with Knowledge Graphs

<https://doi.org/10.1515/bp-2020-0042>

Abstract: The transfer of knowledge has not changed fundamentally for many hundreds of years. It is usually document-based formerly printed on paper as a classic essay and nowadays as PDF. With around 2.5 million new research contributions every year, researchers drown in a flood of pseudo-digitized PDF publications. As a result research is seriously weakened. In this article, we argue for representing scholarly contributions in a structured and semantic way as a knowledge graph. The advantage is that information represented in a knowledge graph is readable by machines and humans. As an example, we give an overview on the Open Research Knowledge Graph (ORKG), a service implementing this approach. For creating the knowledge graph representation, we rely on a mixture of manual (crowd/expert sourcing) and (semi-)automated techniques. Only with such a combination of human and machine intelligence, we can achieve the required quality of the representation to allow for novel exploration and assistance services for researchers. As a result, a scholarly knowledge graph such as the ORKG can be used to give a condensed overview on the state-of-the-art addressing a particular research quest, for example as a tabular comparison of contributions according to various characteristics of the approaches. Further possible intuitive access interfaces to such scholarly knowledge graphs include domain-specific (chart) visualizations or answering of natural language questions.

*Corresponding author: Prof. Dr. Sören Auer, soeren@tib.tu-berlin.de
Allard Oelen, allard.oelen@tib.tu-berlin.de
Muhammad Haris, muhhammad.haris@tib.tu-berlin.de
Dr. Markus Stocker, markus.stocker@tib.tu-berlin.de
Dr. Jennifer D'Souza, jennifer.dsouza@tib.tu-berlin.de
Kheir Eddine Farfar, kheir.farfar@tib.tu-berlin.de
Lars Vogt, lars.vogt@tib.tu-berlin.de
Manuel Prinz, manuel.prinz@tib.tu-berlin.de
Vitalis Wiens, vitalis.wiens@tib.tu-berlin.de
Mohamad Yaser Jaradeh, yaser.jaradeh@tib.tu-berlin.de

Keywords: Subject classification; knowledge graph; semantic web; crowdsourcing; text mining

Zusammenfassung: Zugang zu wissenschaftlicher Literatur mit Wissensgraphen

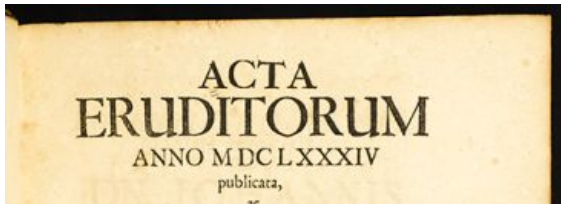
Zusammenfassung: Der Verbreitung wissenschaftlicher Erkenntnisse hat sich seit vielen hundert Jahren nicht grundlegend verändert: Er erfolgt in der Regel dokumentenbasiert – früher als klassischer Aufsatz auf Papier gedruckt und heute online als PDF. Mit rund 2,5 Millionen neuen Forschungsbeiträgen pro Jahr ertrinken Forscher in einer Flut von pseudo-digitalisierten PDF-Publikationen. Als Folge davon wird die Forschung stark geschwächt. In diesem Artikel plädieren wir dafür, wissenschaftliche Beiträge in strukturierter und semantischer Form als Wissensgraph zu repräsentieren. Der Vorteil ist, dass die in einem Wissensgraph dargestellten Informationen für Maschinen und Menschen lesbar sind. Als Beispiel geben wir einen Überblick über den Open Research Knowledge Graph (ORKG), einen Dienst, der diesen Ansatz umsetzt. Für die Erstellung des Wissensgraph setzen wir eine Mischung aus manuellen (crowd/expert sourcing) und (halb-)automatisierten Techniken ein. Nur mit einer solchen Kombination aus menschlicher und maschineller Intelligenz können wir die erforderliche Qualität der Darstellung erreichen, um neuartige Explorations- und Unterstützungsdienste für Forscher zu ermöglichen. In Ergebnis kann ein Wissensgraph wie der ORKG verwendet werden, um einen kompakteren Überblick über den Stand der Technik in Bezug auf eine bestimmte Forschungsaufgabe zu geben, z. B. als tabellarischer Vergleich der Beiträge nach verschiedenen Merkmalen der Ansätze. Weitere mögliche intuitive Nutzungsoberflächen zu solchen wissenschaftlichen Wissensgraphen sind domänen-spezifische Visualisierungen oder die Beantwortung natürlichsprachlicher Fragen mittels Question Answering.

Schlüsselwörter: Sächerschließung; Wissensgraph; Semantic Web; Crowdsourcing; Text Mining

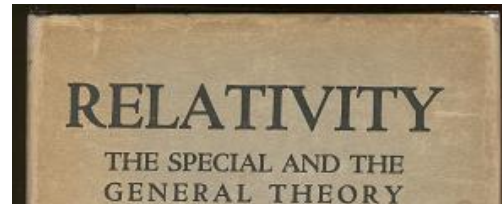
Not much has changed!

What about Science?

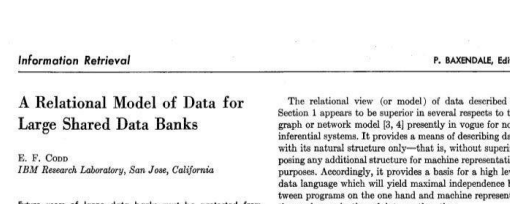
Over 300 years ago



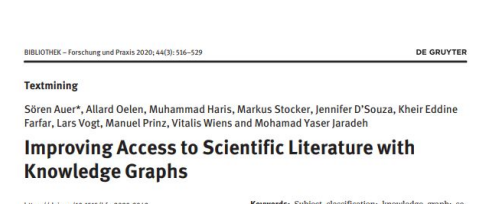
100 years ago



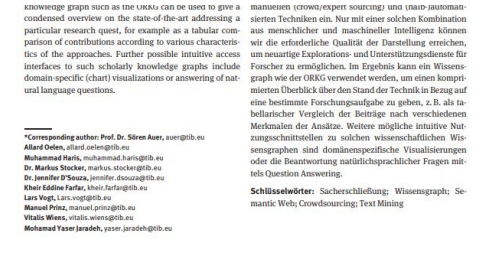
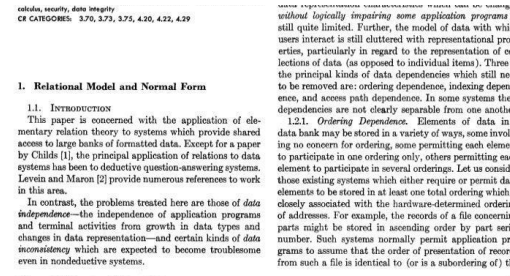
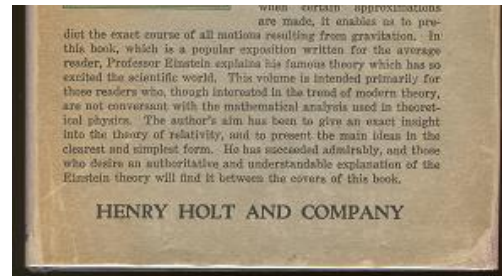
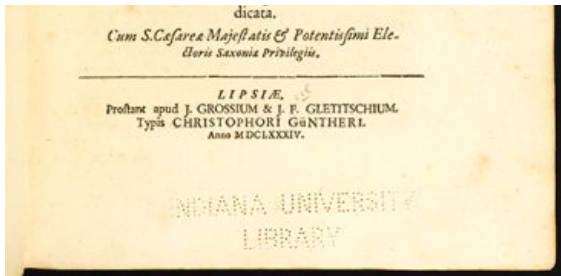
20 years ago



Today



Science does not harvest the full potential of digitalization



Not much has changed!

A Consequence of Document Centered Information Flows: The Publication Flood

- ~ 2.5 Mio new publications per year
- Researchers lack overview, even in small fields
- Loss of knowledge
- Answering questions is like looking for a needle in the haystack




An Example – CRISPR



[HTML] The heroes of CRISPR

ES Lander - Cell, 2016 - Elsevier


... for CRISPR-based resistance, they set out to create the first artificial CRISPR arrays—programming CRISPR ... As they predicted, the strains carrying the new CRISPR sequence showed ...

☆ Speichern  Zitieren Zitiert von: 538 Ähnliche Artikel Alle 20 Versionen

A CRISPR view of development

MM Harrison, BV Jenkins... - Genes & ..., 2014 - genesdev.cshlp.org

... as “spacers” between repetitive sequences in the CRISPR locus of the host genome. The CRISPR locus is transcribed and processed into short CRISPR RNAs (crRNAs) that guide the ...

☆ Speichern  Zitieren Zitiert von: 272 Ähnliche Artikel Alle 10 Versionen

[HTML] CRISPR-based diagnostics

MM Kaminski, OO Abudayyeh, JS Gootenberg... - Nature Biomedical ..., 2021 - nature.com

... with the CRISPR-associated (Cas) enzyme. Although there are diverse CRISPR–Cas ... these systems are connected by their dependence on CRISPR RNA (crRNA), which guides ...

☆ Speichern  Zitieren Zitiert von: 59 Ähnliche Artikel Alle 10 Versionen

-
-
-

Specific research questions:

- Who applied CRISPR to butterflies?
- How to apply CRISPR with minimal costs?
- How do different genome editing techniques compare?

The Publication Flood – More than just an Inconvenience for Scientists

- Globally almost \$1,700,000,000,000 (1.7 trillion) spent on research & development
 - Large share wasted in inefficient system
- Costs time & money!



Further Challenges of Document-Oriented



Reproducibility Crisis



ELSEVIER

Monopolization of
commercial actors



Deficiency of
Peer-Review



Lack of machine assistance



Predatory Publishing

Time to Rethink Scholarly Communication!

The solution is not „better pdfs“...



*“The lightbulb was **not** invented by improving the candle.”*

Oren Harari

Digitalization is **more** than just Digitization!
Current and future scientific challenges can not be tackled with an outdated communication system.

**Digitalize Knowledge,
Not Documents!**

The Open Research Knowledge Graph



ORKG

As the name already suggests, ORKG is a **knowledge graph**.

Knowledge Graphs are widely used in industry...



Why not use them for (open) science as well?

Knowledge Graphs are widely used in industry...



Fine, but what is such a knowledge graph?

Why not use them for (open) science as well?

Representation of Information



There is a lot of information in a text...

bioRxiv posts many COVID19-related papers. A reminder: they have not been formally peer-reviewed and should not guide health-related behavior or be reported in the press as conclusive.

New Results

Follow this preprint

A practical guide to CRISPR/Cas9 genome editing in Lepidoptera

Linlin Zhang, Robert D. Reed

doi: <https://doi.org/10.1101/130344>

Now published in *Diversity and Evolution of Butterfly Wing Patterns* doi: 10.1007/978-981-10-4956-9_8



Abstract Full Text Info/History Metrics

Review PDF

Abstract

CRISPR/Cas9 genome editing has revolutionized functional genetic work in many organisms and is having an especially strong impact in emerging model systems. Here we summarize recent advances in applying CRISPR/Cas9 methods in Lepidoptera, with a focus on providing practical advice on the entire process of genome editing from experimental design through to genotyping. We also describe successful targeted GFP knockins that we have achieved in butterflies. Finally, we provide a complete, detailed protocol for producing targeted long deletions in butterflies.

- Metadata
- Research problem
- Methods
- Material
- Results
- ...

Representation of Information



bioRxiv posts many COVID19-related papers. A reminder: they have not been formally peer-reviewed and should not guide health-related behavior or be reported in the press as conclusive.

New Results [Follow this preprint](#)

A practical guide to CRISPR/Cas9 genome editing in Lepidoptera

Linlin Zhang, Robert D. Reed

doi: <https://doi.org/10.1101/130344>

Now published in *Diversity and Evolution of Butterfly Wing Patterns* doi: [10.1007/978-981-10-4956-9_8](https://doi.org/10.1007/978-981-10-4956-9_8)



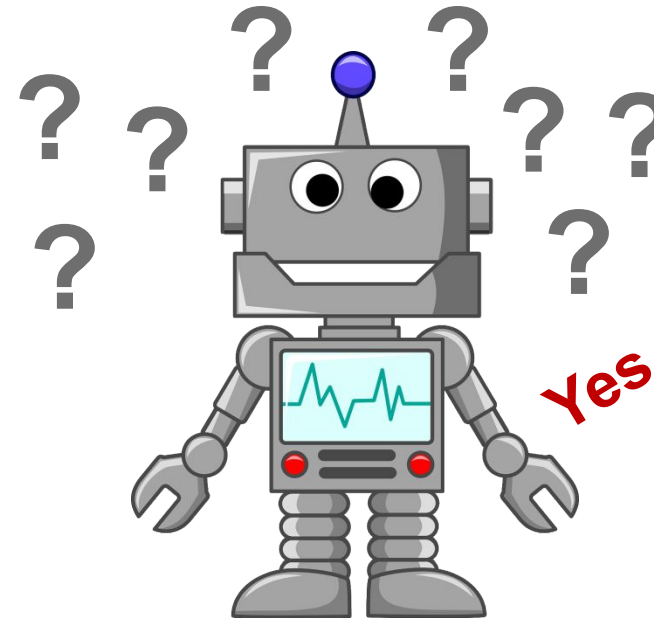
Abstract Full Text Info/History Metrics [Preview PDF](#)

Abstract

CRISPR/Cas9 genome editing has revolutionized functional genetic work in many organisms and is having an especially strong impact in emerging model systems. Here we summarize recent advances in applying CRISPR/Cas9 methods in Lepidoptera, with a focus on providing practical advice on the entire process of genome editing from experimental design through to genotyping. We also describe successful targeted GFP knockins that we have achieved in butterflies. Finally, we provide a complete, detailed protocol for producing targeted long deletions in butterflies.

There is a lot of information in a text...

...that can unfortunately not be understood by a machine.

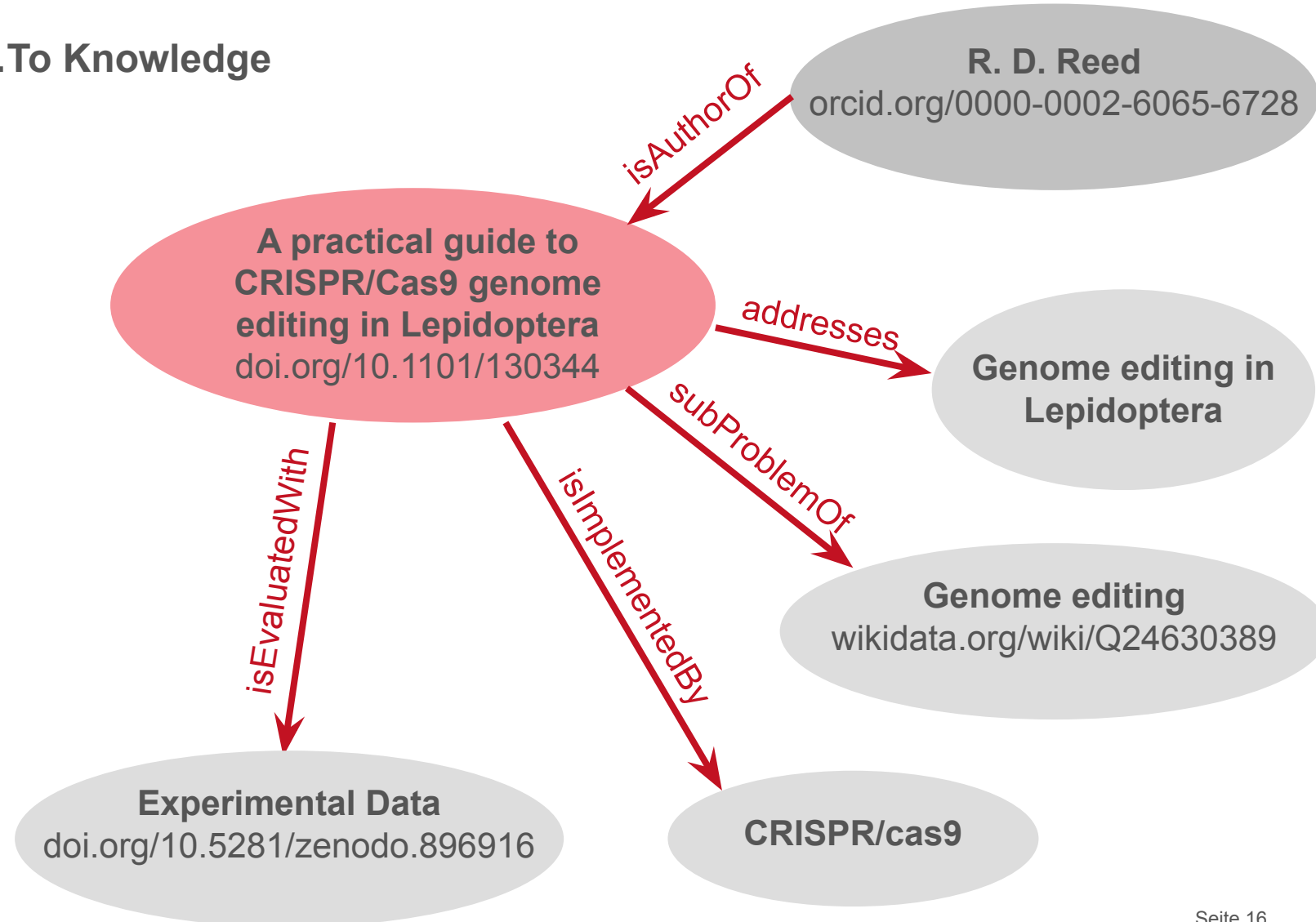
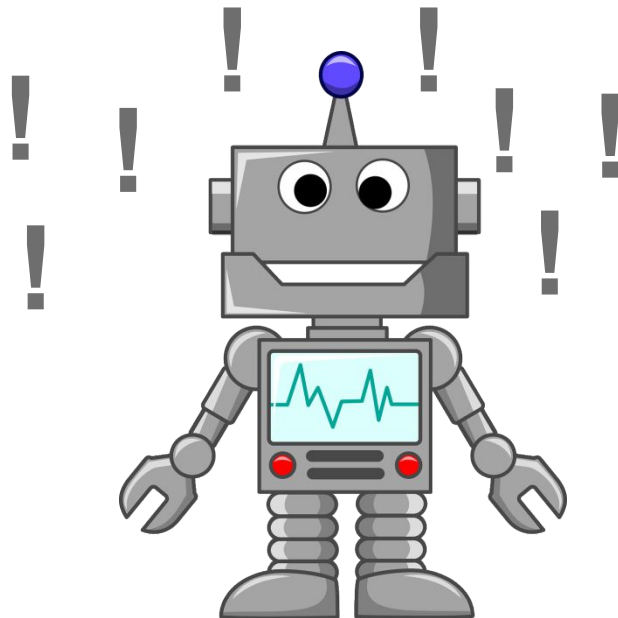


Yes, including ChatGPT

Knowledge Representation in Graphs

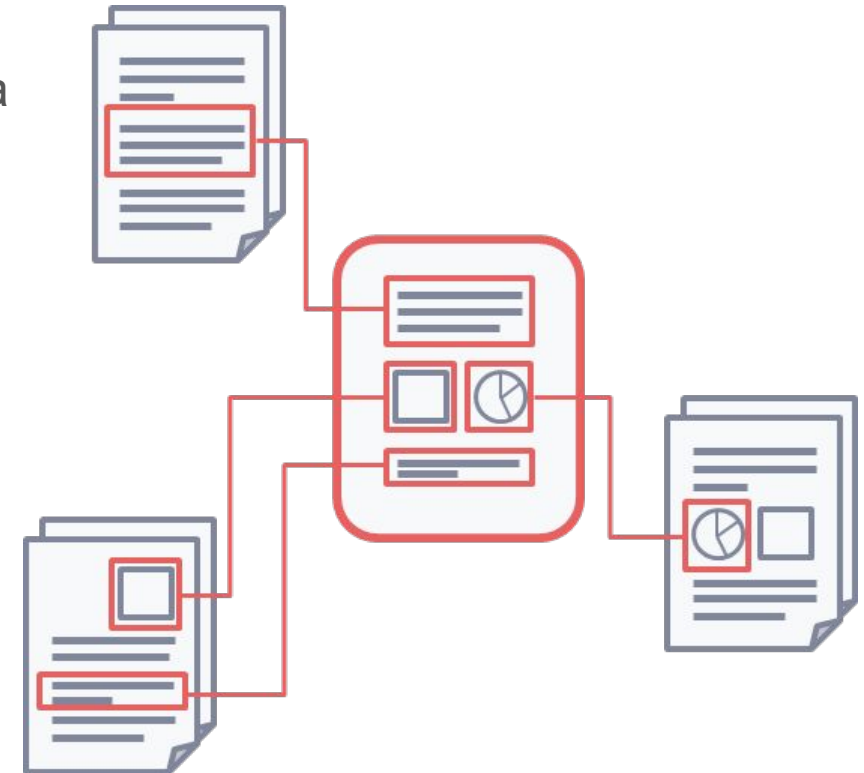
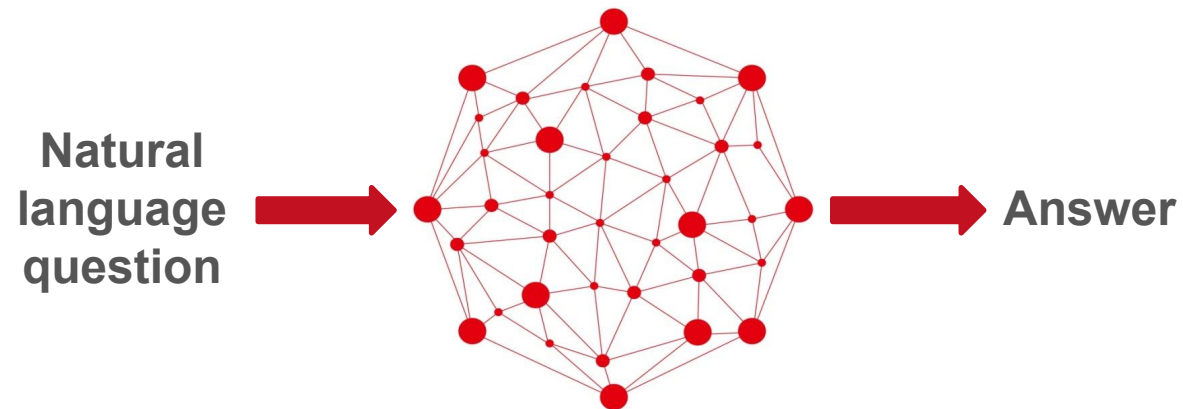
From papers...

...To Knowledge



Advantages of a Graph-Based Approach

- Machine-actionable
- Automated finding and linking of research contributions towards a specific problem
- Natural language question answering possible
e.g. „How do different genome editing techniques compare?“



- Explore knowledge in entirely new ways

An Example: SARS-CoV 2 Basic Reproduction Number



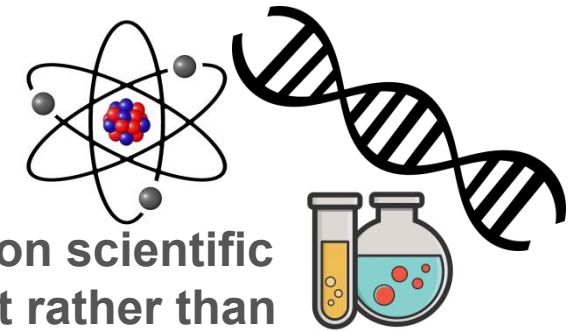
Properties	The early phase of the COVID-19 outbreak in Lombardy, Italy 2020 - Contribution 1	Transmission potential of COVID-19 in Iran 2020 - Contribution 1	Transmission potential of COVID-19 in Iran 2020 - Contribution 2	Estimating the generation interval for COVID-19 based on symptom onset data 2020 - Contribution 1
location	Lombardy, Italy	Iran	Iran	Singapore
Time period	Time interval	Time interval	Time interval	Time interval
has beginning	2020-01-14	2020-02-19	2020-02-19	2020-01-21
has end	2020-03-08	2020-02-29	2020-02-29	2020-02-26
Basic reproduction number	Basic reproduction number estimate value specification	Basic reproduction number estimate value specification	Basic reproduction number estimate value specification	Basic reproduction number estimate value specification
Has value	3.1	3.6	3.58	1.27
Confidence interval (95%)	Confidence interval (95%)	Confidence interval (95%)	Confidence interval (95%)	Confidence interval (95%)
Lower confidence limit	2.9	3.4	1.29	1.19
Upper confidence limit	3.2	4.2	8.46	1.36
Method*		generalized growth model	based on the calculation of the epidemic's doubling times: estimated epidemic doubling time of 1.20 (95% CI, 1.05, 1.44) days	generation interval

ORKG's Objectives

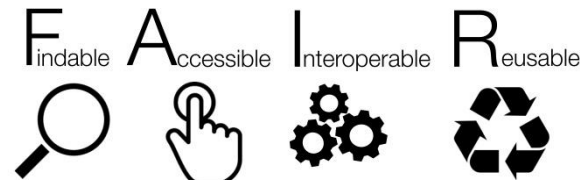


Provide overview over the state-of-the-art for specific research problems

Foster collaboration



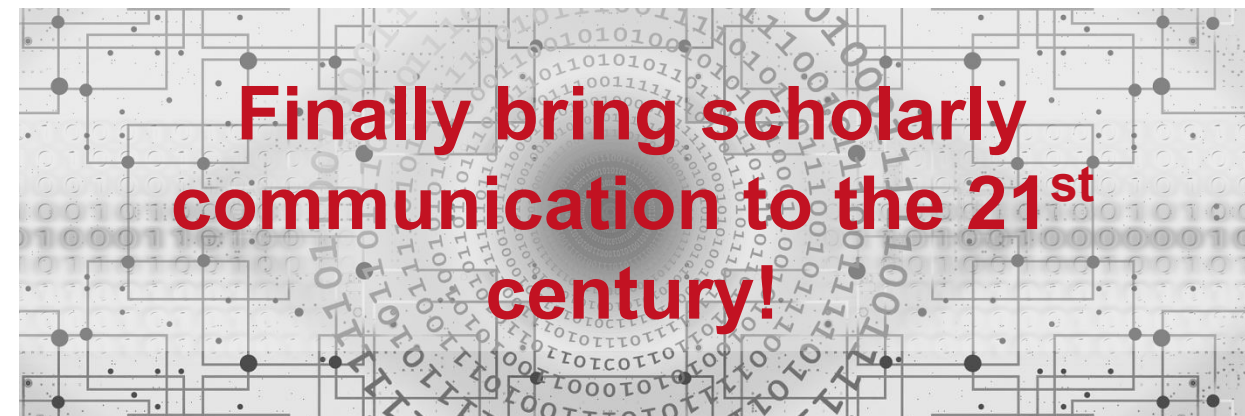
Focus on scientific content rather than document



Make research FAIR

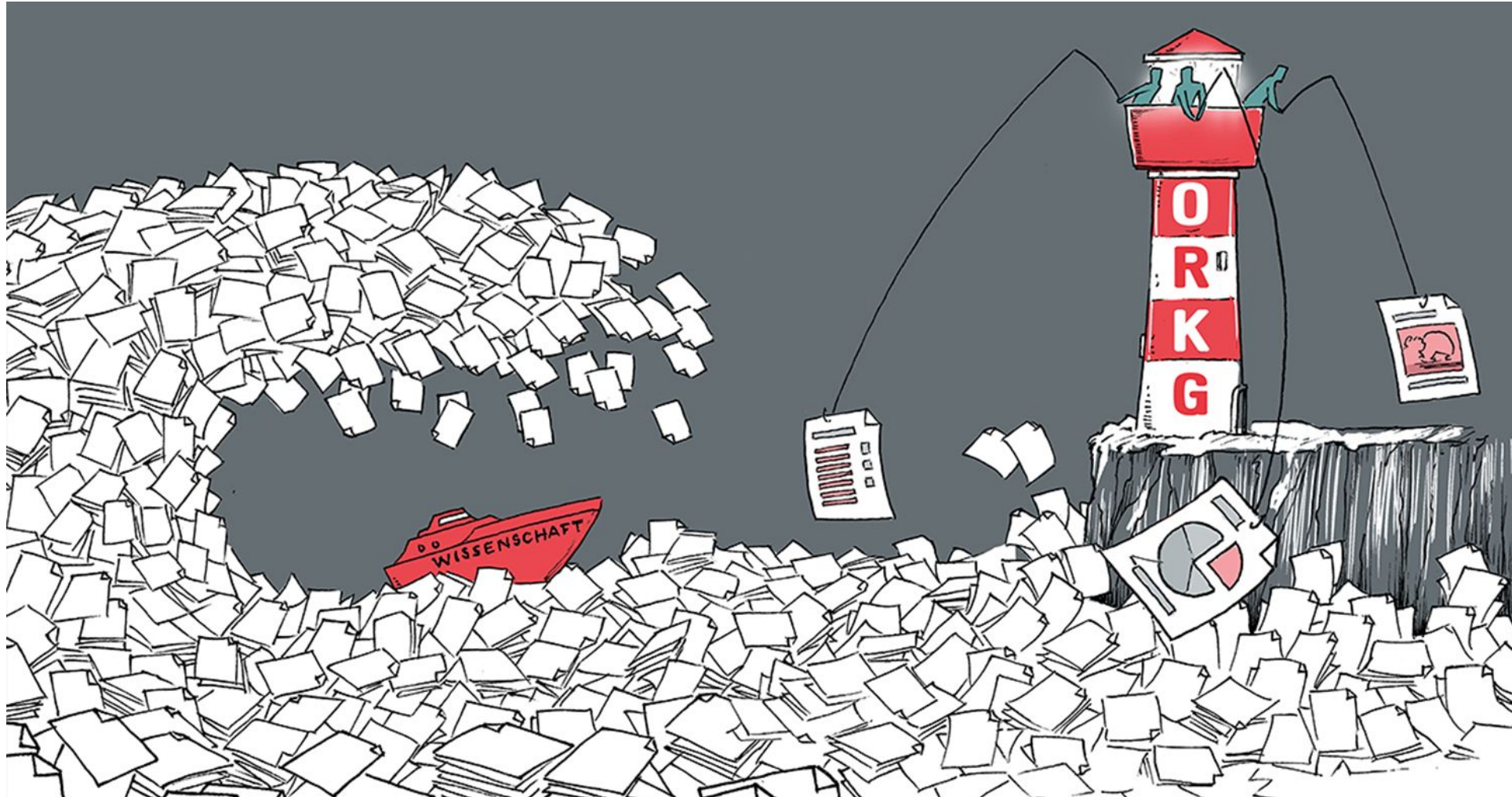


Tackle interdisciplinary challenges such as climate change research, disease prevention, etc.



Finally bring scholarly communication to the 21st century!

ORKG: Lighthouse in the Publication Flood



View ▾

Tools ▾

About ▾

Comparisons

Papers

Visualizations

Reviews Beta

Lists Beta

Benchmarks

**What can you do
with the ORKG?**

ORKG, papers are easier to

**Let's have a look
at the content!**

Current Status

- ~ 14.500 Papers described
- ~ 1100 Comparisons
- ~ 5.000 Research questions/ problems
- ~ 1200 Users
- ~ 30 Organizations

...could be more!

So how do we get more content?

Who creates ORKG content?



Translation



bioRxiv posts many COVID19-related papers. not guide health-related behavior or be reported

New Results

A practical guide to CRISPR/Cas9

Linlin Zhang, Robert D. Reed

doi: <https://doi.org/10.1101/130344>

Now published in *Diversity and Evolution of Life*



Abstract Full Text Info/History

Abstract

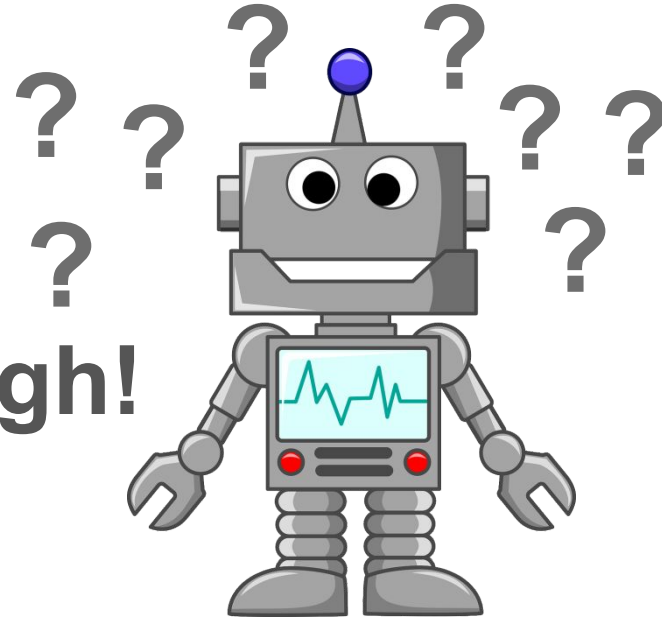
CRISPR/Cas9 genome editing has revolutionized genetics and is having an especially strong impact on the study of recent advances in applying CRISPR/Cas9 to genome editing. We provide practical advice on the entire process from target identification to genotyping. We also describe successful applications in butterflies. Finally, we provide a complete protocol for genome editing in butterflies.



R. D. Reed
/0000-0002-6065-6728

Machines?

Not precise enough!



Genome editing in
Lepidoptera

Genome editing
[a.org/wiki/Q24630389](https://www.wikidata.org/wiki/Q24630389)

Experimental Data
doi.org/10.5281/zenodo.896916

CRISPR/cas9

Who creates ORKG content?



Translation



bioRxiv posts many COVID19-related papers. not guide health-related behavior or be reported

New Results

A practical guide to CRISPR/Cas9

Linlin Zhang, Robert D. Reed

doi: <https://doi.org/10.1101/130344>

Now published in *Diversity and Evolution of Life*

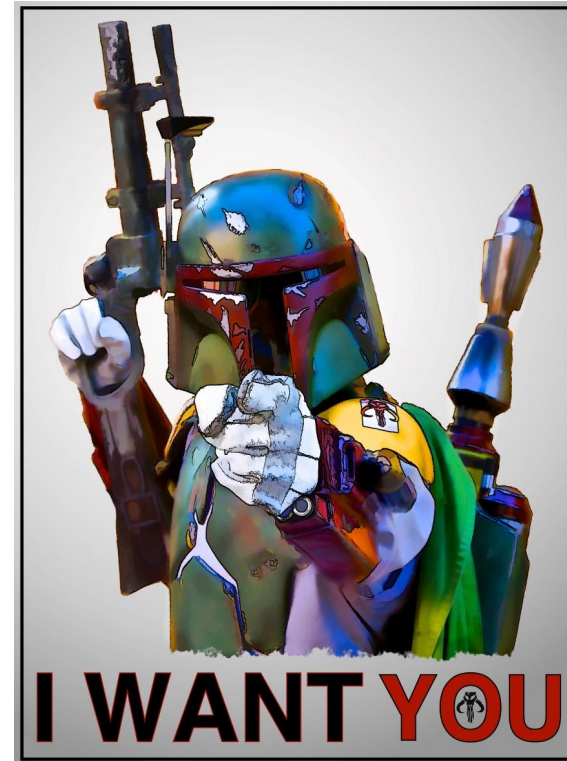


Abstract Full Text Info/History

Abstract

CRISPR/Cas9 genome editing has revolutionized genetics and is having an especially strong impact on developmental biology. Recent advances in applying CRISPR/Cas9 to genome editing provide practical advice on the entire process from target identification to genotyping. We also describe success stories in genome editing of butterflies. Finally, we provide a complete protocol for genome editing in butterflies.

Better: Scientific Communities!



R. D. Reed
reed@tib.tu-berlin.de
+49 30 7000-0002-6065-6728

Genome editing in
Lepidoptera

Genome editing
www.wikidata.org/wiki/Q24630389

Experimental Data
doi.org/10.5281/zenodo.896916

CRISPR/cas9

Who creates ORKG content?

Crowd-based approach for the curation process

Following the principle of Wikipedia:
Everyone can create, edit, add, complement, reuse, etc.



How to get out the most of ORKG for your discipline?

Content

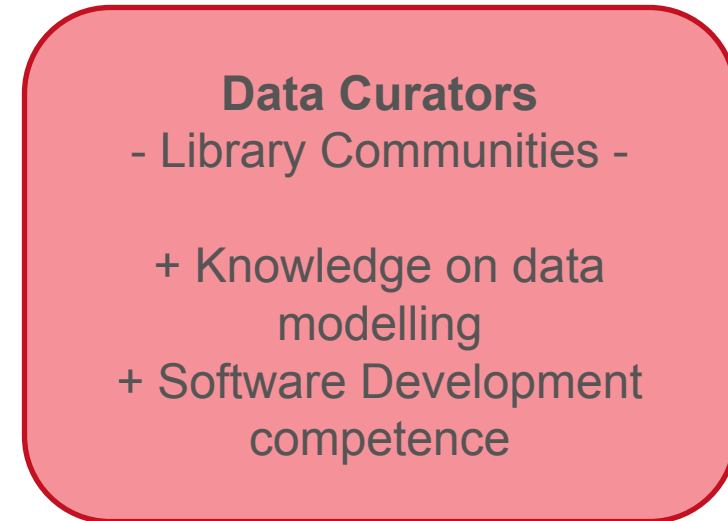


Features

Requirements



ORKG Curation – Different Expertise

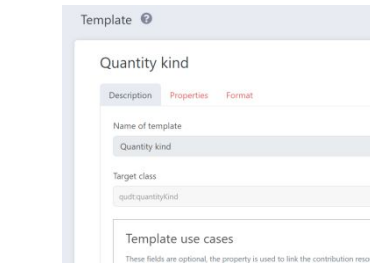


Observatories: Taking the Lead in Content Curation



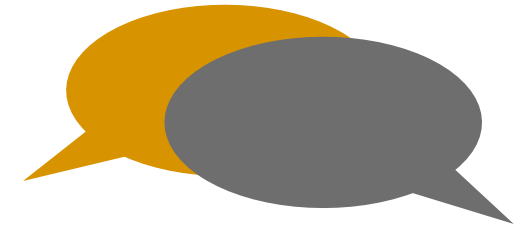
Organize research
in your field

Ensure high
quality standard



Create templates
and simplify using
ORKG for beginners

Promote ORKG



Stay in contact with
development team:
Issues & Requests will be
prioritized

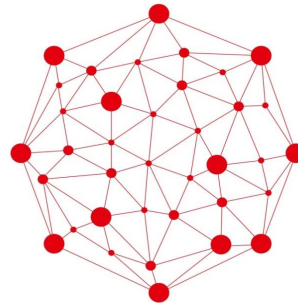


Summary

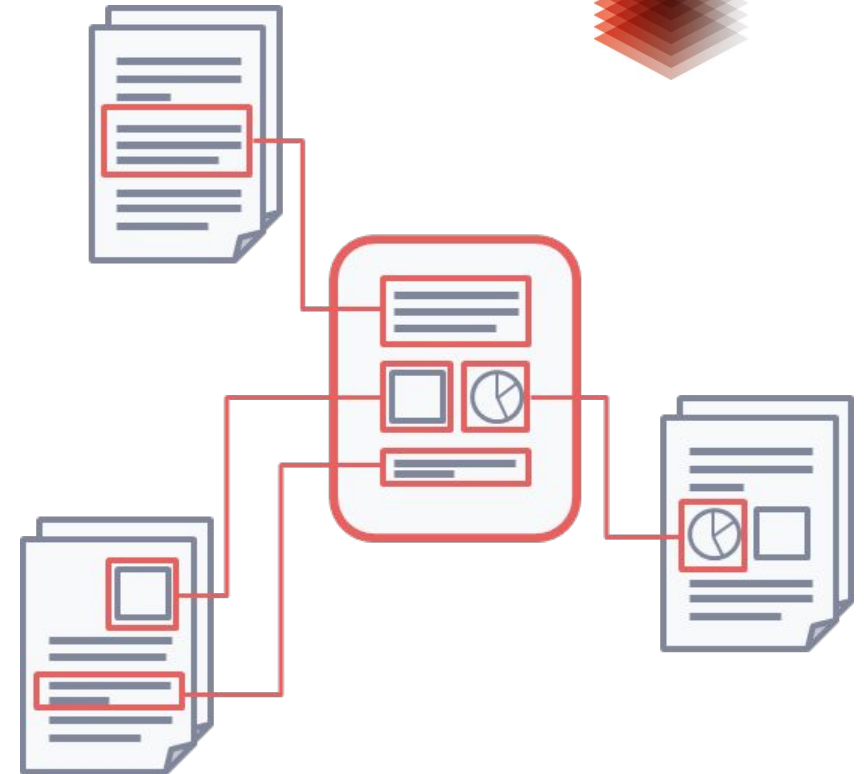


Rethink
scholarly
communication

Machine-actionable
knowledge representation



Crowd-based
approach



Learn more: orkg.org
Contact us: info@orkg.org
Follow us: [@orkg_org](https://twitter.com/orkg_org)