LEIBNIZ INFORMATION CENTRE
FOR SCIENCE AND TECHNOLOGY
UNIVERSITY LIBRARY

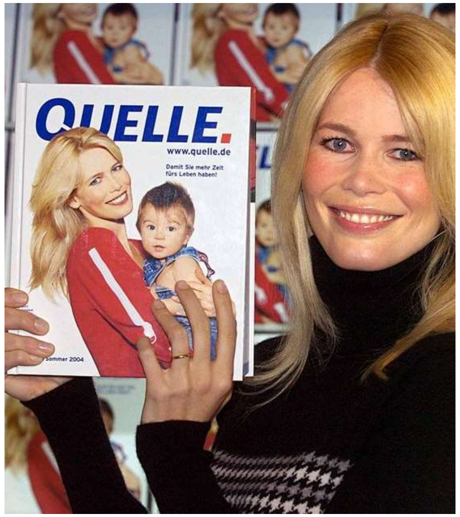
TIB

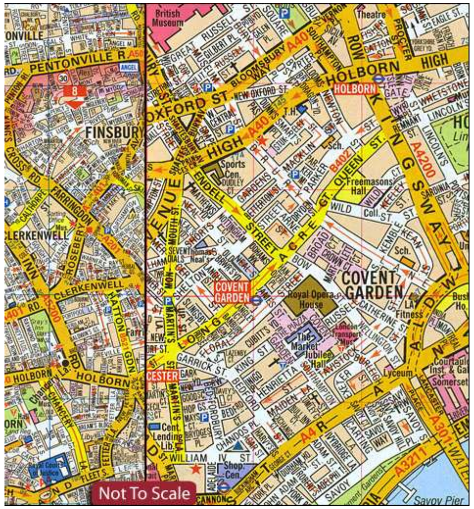
Leibniz Universität Hannover

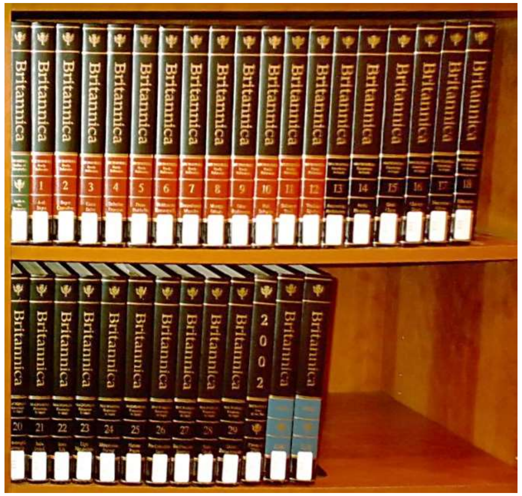
Von dokumentenbasierten zu wissensbasierten Informationsflüssen – Die wissenschaftlichen Bibliotheken im Transformationsprozess

Prof. Dr. Sören Auer
Leibniz University of Hannover
TIB Technische Informationsbibliothek

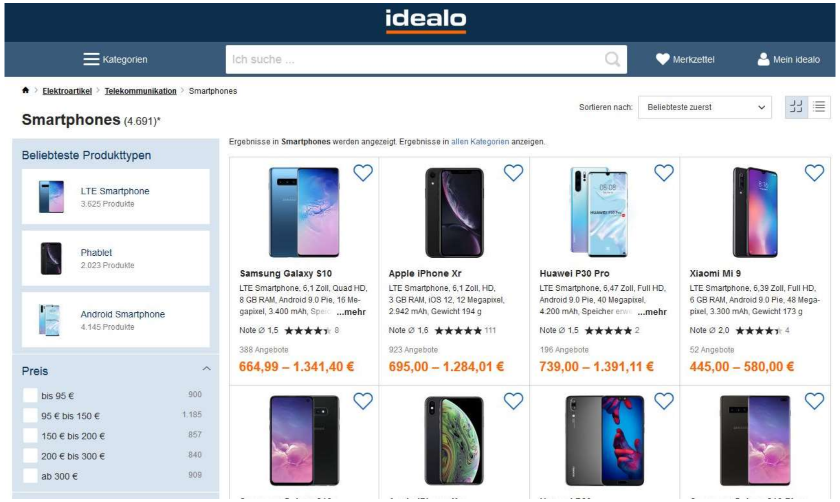
How did information flows change in the digital era?

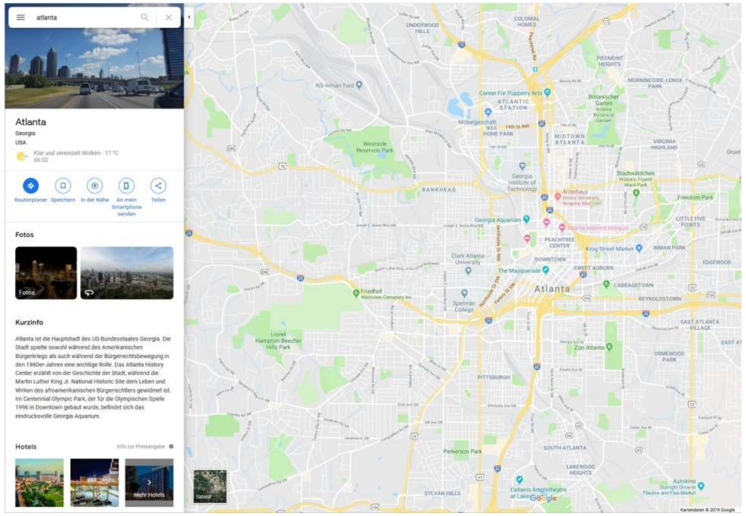


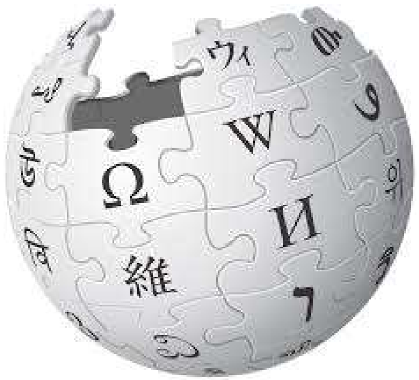


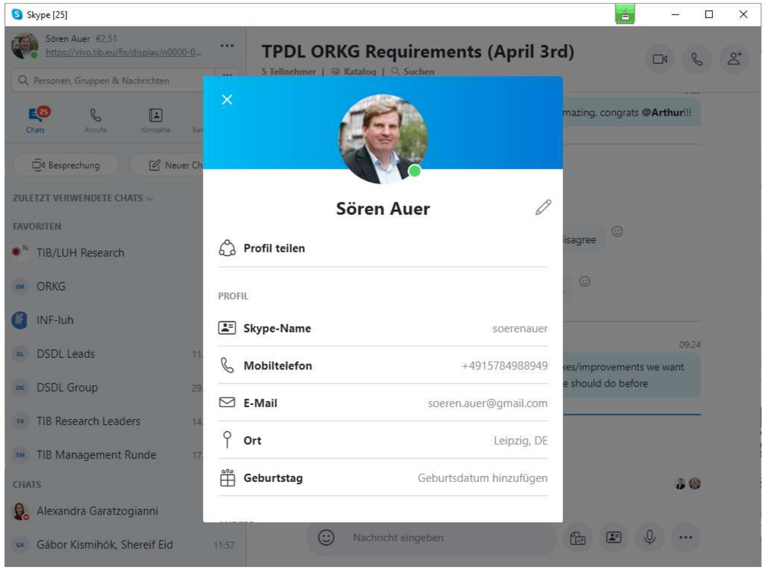


How does it work today?







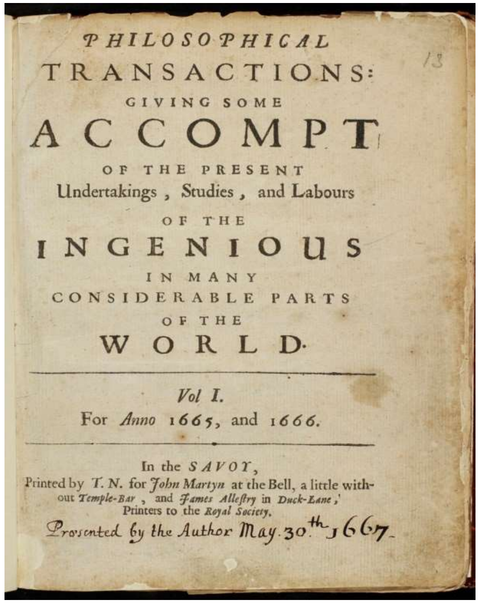


The World of Publishing & Communication has profundely changed

- **New means adapted to the new possibilities** were developed, e.g. „zooming“, dynamics
- **Business models** changed completely
- More focus on data, interlinking of **data / services and search** in the data
- Integration, **crowdsourcing, data curation** play an important role

Scholarly Communication has not changed (much)

17th century



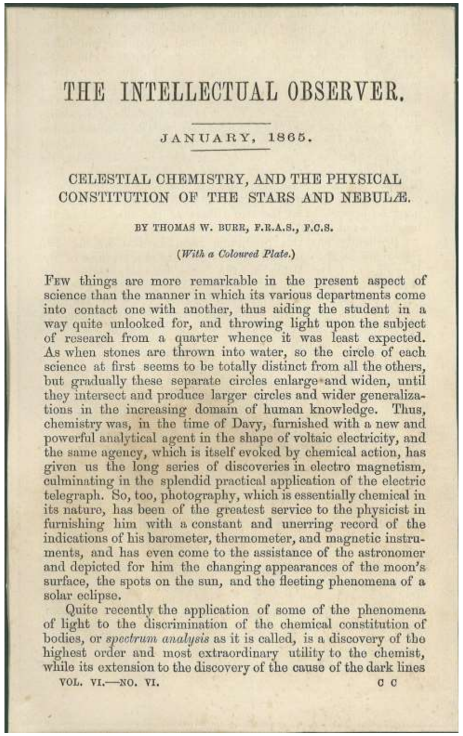
PHILOSOPHICAL TRANSACTIONS:
GIVING SOME
ACCOMPT
OF THE PRESENT
Undertakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD.

Vol I.
For Anno 1665, and 1666.

In the SAVOY,
Printed by T. N. for John Martyn at the Bell, a little without Temple-Bar, and James Allestry in Duck-Lane, Printers to the Royal Society.

Presented by the Author May 30th 1667

19th century



THE INTELLECTUAL OBSERVER.

JANUARY, 1865.

CELESTIAL CHEMISTRY, AND THE PHYSICAL CONSTITUTION OF THE STARS AND NEBULÆ.

BY THOMAS W. BURR, F.R.A.S., F.C.S.

(With a Coloured Plate.)

Few things are more remarkable in the present aspect of science than the manner in which its various departments come into contact one with another, thus aiding the student in a way quite unlooked for, and throwing light upon the subject of research from a quarter whence it was least expected. As when stones are thrown into water, so the circle of each science at first seems to be totally distinct from all the others, but gradually these separate circles enlarge and widen, until they intersect and produce larger circles and wider generalizations in the increasing domain of human knowledge. Thus, chemistry was, in the time of Davy, furnished with a new and powerful analytical agent in the shape of voltaic electricity, and the same agency, which is itself evoked by chemical action, has given us the long series of discoveries in electro magnetism, culminating in the splendid practical application of the electric telegraph. So, too, photography, which is essentially chemical in its nature, has been of the greatest service to the physicist in furnishing him with a constant and unerring record of the indications of his barometer, thermometer, and magnetic instruments, and has even come to the assistance of the astronomer and depicted for him the changing appearances of the moon's surface, the spots on the sun, and the fleeting phenomena of a solar eclipse.

Quite recently the application of some of the phenomena of light to the discrimination of the chemical constitution of bodies, or *spectrum analysis* as it is called, is a discovery of the highest order and most extraordinary utility to the chemist, while its extension to the discovery of the cause of the dark lines

VOL. VI.—NO. VI. C C

20th century

Information Retrieval P. BAXENDALE, Editor

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 types of stored 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 bank, 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
CR CATEGORIES: 3.70, 3.73, 3.75, 4.20, 4.22, 4.29

1. Relational Model and Normal Form

1.1. Introduction

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

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 noninferential 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 2 on the "connection trap").

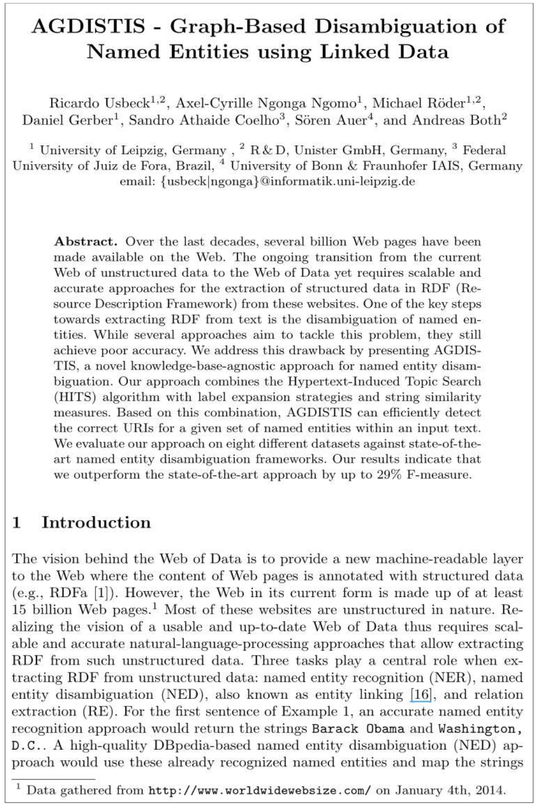
Finally, the relational view permits a clearer evaluation of the scope and logical limitations of present formatted data systems, and also the relative merits (from a logical standpoint) of competing representations of data within a single system. Examples of this clearer perspective are cited in various parts of this paper. Implementations of systems to support the relational model are not discussed.

1.2. Data Dependencies in Present 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

21st century



AGDISTIS - Graph-Based Disambiguation of Named Entities using Linked Data

Ricardo Usbeck[1,2], Axel-Cyrille Ngonga Ngomo[1], Michael Röder[1,2], Daniel Gerber[1], Sandro Athaide Coelho[3], Sören Auer[4], and Andreas Both[2]

[1] University of Leipzig, Germany, [2] R & D, Unister GmbH, Germany, [3] Federal University of Juiz de Fora, Brazil, [4] University of Bonn & Fraunhofer IAIS, Germany
email: {usbeck|ngonga}@informatik.uni-leipzig.de

Abstract. Over the last decades, several billion Web pages have been made available on the Web. The ongoing transition from the current Web of unstructured data to the Web of Data yet requires scalable and accurate approaches for the extraction of structured data in RDF (Resource Description Framework) from these websites. One of the key steps towards extracting RDF from text is the disambiguation of named entities. While several approaches aim to tackle this problem, they still achieve poor accuracy. We address this drawback by presenting AGDISTIS, a novel knowledge-base-agnostic approach for named entity disambiguation. Our approach combines the Hypertext-Induced Topic Search (HITS) algorithm with label expansion strategies and string similarity measures. Based on this combination, AGDISTIS can efficiently detect the correct URIs for a given set of named entities within an input text. We evaluate our approach on eight different datasets against state-of-the-art named entity disambiguation frameworks. Our results indicate that we outperform the state-of-the-art approach by up to 29% F-measure.

1 Introduction

The vision behind the Web of Data is to provide a new machine-readable layer to the Web where the content of Web pages is annotated with structured data (e.g., RDFa [1]). However, the Web in its current form is made up of at least 15 billion Web pages.[1] Most of these websites are unstructured in nature. Realizing the vision of a usable and up-to-date Web of Data thus requires scalable and accurate natural-language-processing approaches that allow extracting RDF from such unstructured data. Three tasks play a central role when extracting RDF from unstructured data: named entity recognition (NER), named entity disambiguation (NED), also known as entity linking [16], and relation extraction (RE). For the first sentence of Example 1, an accurate named entity recognition approach would return the strings `Barack Obama` and `Washington, D.C.`. A high-quality DBpedia-based named entity disambiguation (NED) approach would use these already recognized named entities and map the strings

[1] Data gathered from `http://www.worldwidewebsize.com/` on January 4th, 2014.

Meanwhile other information intense domains were completely disrupted: mail order catalogs, street maps, phone books, …

We need to rethink the way how research is represented and communicated

Challenges we are facing:

Digitalisation of Science	Monopolisation by commercial actors	Reproducibility Crisis	Proliferation of publications	Deficiency of Peer Review
▪ Data integration and analysis ▪ Digital collaboration	▪ Publisher look-in effects ▪ Maximization of profits [1]	▪ Majority of experiments are hard or not reproducible [2]	▪ Publication output doubled within a decade ▪ continues to rise [3]	▪ Deteriorating quality [4] ▪ Predatory publishing

[1] http://thecostofknowledge.com, https://www.projekt-deal.de
[2] M. Baker: *1,500 scientists lift the lid on reproducibility*, *Nature,* 2016.
[3] *Science and Engineering Publication Output Trends*, National Science Foundation, 2018.
[4] J. Couzin-Frankel: *Secretive and Subjective, Peer Review Proves Resistant to Study.* Science, 2013.

Root Cause –
Deficiency of Scholarly Communication?

Lack of…

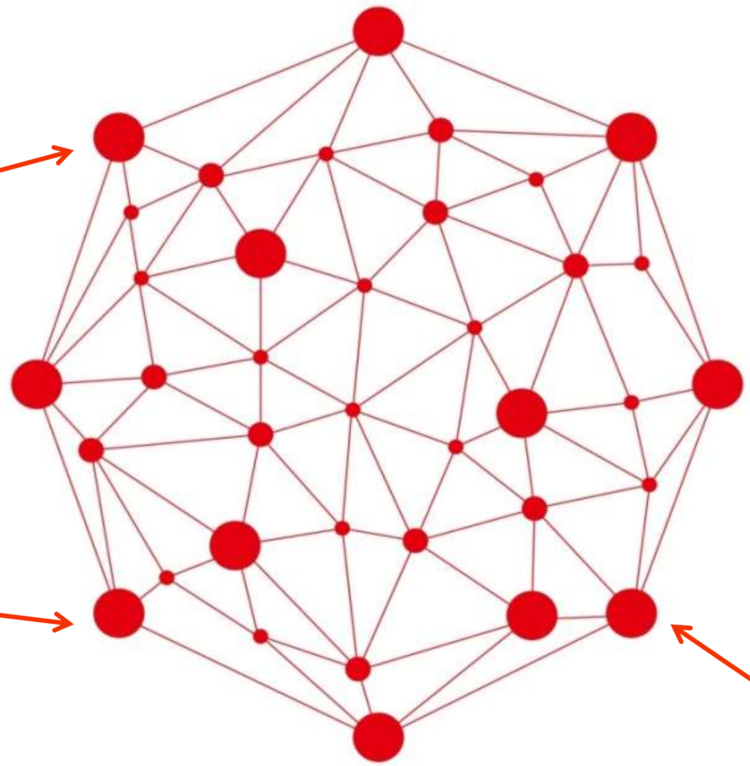
Transparency	Integratability	Machine assistance
information is hidden in text	fitting different research results together	unstructured content is hard to process

Identifyability	Collaboration	Overview
of concepts beyond metadata	one brain barrier	Schientists look for the needle in the haystack

How can
we fix it?

TIB

Concepts



Overarching Concepts

- Research problems
- Definitions
- Research approaches
- Methods

Artefacts

- Publications
- Data
- Software
- Image/Audio/Video
- Knowledge Graphs / Ontologies

Domain specific Concepts

Mathematics

- Definitions
- Theorems
- Proofs
- Methods
- …

Physics

- Experiments
- Data
- Models
- …

Chemistry

- Substances
- Structures
- Reactions
- …

Computer Science

- Concepts
- Implemen-tations
- Evaluations
- …

Technology

- Standards
- Processes
- Elements
- Units, Sensor data

Architecture

- Regulations
- Elements
- Models
- …

Chemistry Example: CRISPR Genome Editing

TIB

CSH Cold Spring Harbor Laboratory

bioRχiv

THE PREPRINT SERVER FOR BIOLOGY

HOME | ABOUT | SUBMIT | ALERTS / RSS | CHANNELS

Search

Advanced Search

New Results

Previous Next

A practical guide to CRISPR/Cas9 genome editing in Lepidoptera

Posted June 22, 2017.

Linlin Zhang, Robert 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

Download PDF | Share

Email | Citation Tools

Tweet | Like 0 | G+

Abstract | Info/History | Metrics | Preview PDF

Subject Area

Genetics

Subject Areas

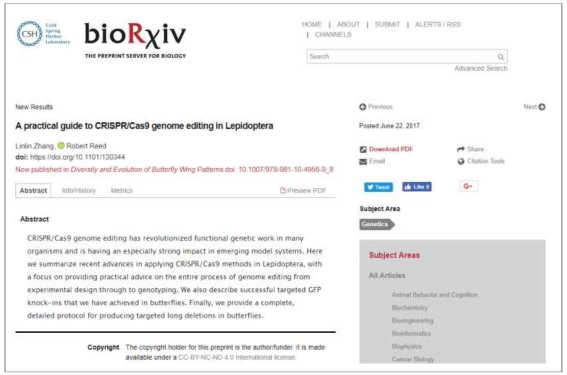
All Articles

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

Source: https://cacm.acm.org/system/assets/0002/2618/021116_Google_KnowledgeGraph.large.jpg?1476779500&1455222197

Chemistry Example: Populating the Graph


bioRxiv
THE PREPRINT SERVER FOR BIOLOGY
HOME | ABOUT | SUBMIT | ALERTS / RSS | CHANNELS
Search
Advanced Search

New Results

A practical guide to CRISPR/Cas9 genome editing in Lepidoptera

Linlin Zhang, Robert 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 | 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 knock-ins that we have achieved in butterflies. Finally, we provide a complete, detailed protocol for producing targeted long deletions in butterflies.

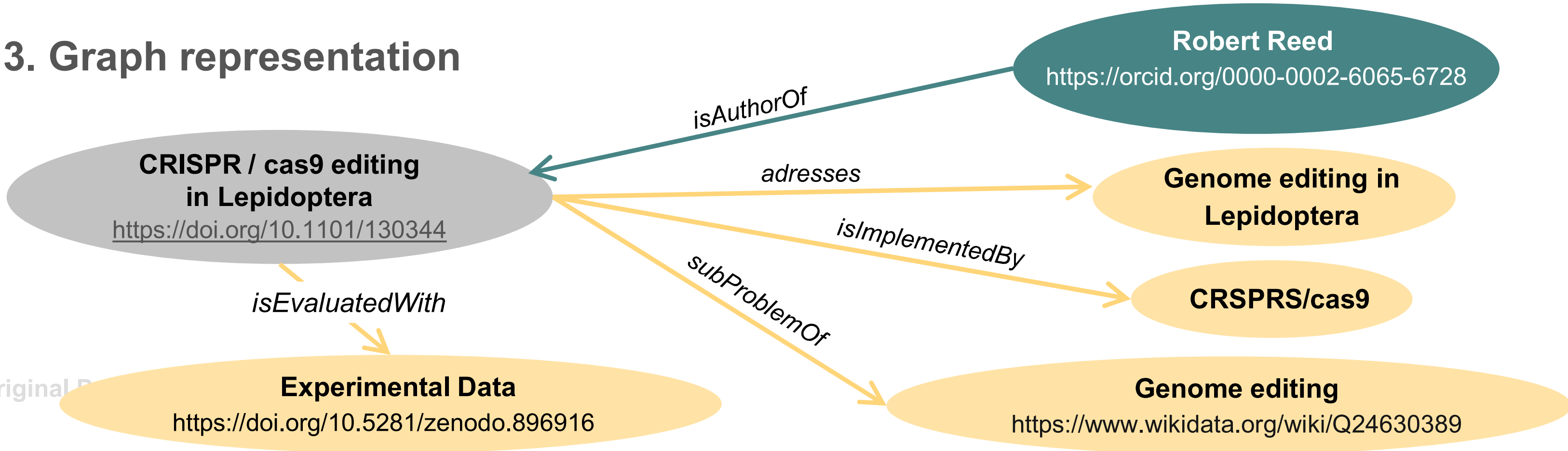
Copyright The copyright holder for this preprint is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

Previous | Next
Posted June 22, 2017.
Download PDF | Share
Email | Citation Tools
Subject Area
Genetics
Subject Areas
All Articles
Animal Behavior and Cognition
Biochemistry
Bioengineering
Bioinformatics
Biophysics
Cancer Biology

2. Adaptive Graph Curation & Completion

Author	Robert Reed
Research Problem	Genome editing in Lepidoptera
Methods	CRISPR / cas9
Applied on	Lepidoptera
Experimental Data	https://doi.org/10.5281/zenodo.896916

3. Graph representation

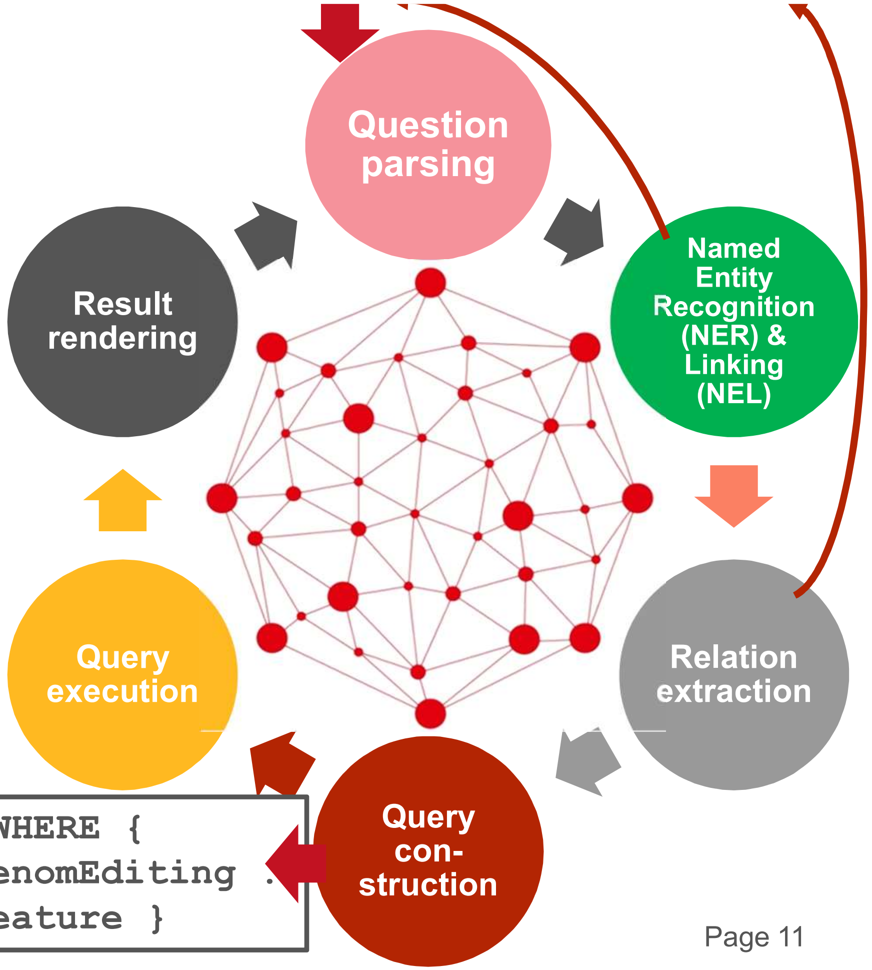


1. Original

Exploration and Question Answering

Research Challenge:
- Intuitive exploration leveraging the rich semantic representations
- Answer natural language questions

Q: *How do different* ***genome editing*** *techniques* ***compare****?*




```
SELECT Approach, Feature WHERE {
  Approach  adresses    GenomEditing .
  Approach  hasFeature  Feature }
```

Result:
Automatic Generation of Comparisons / Surveys

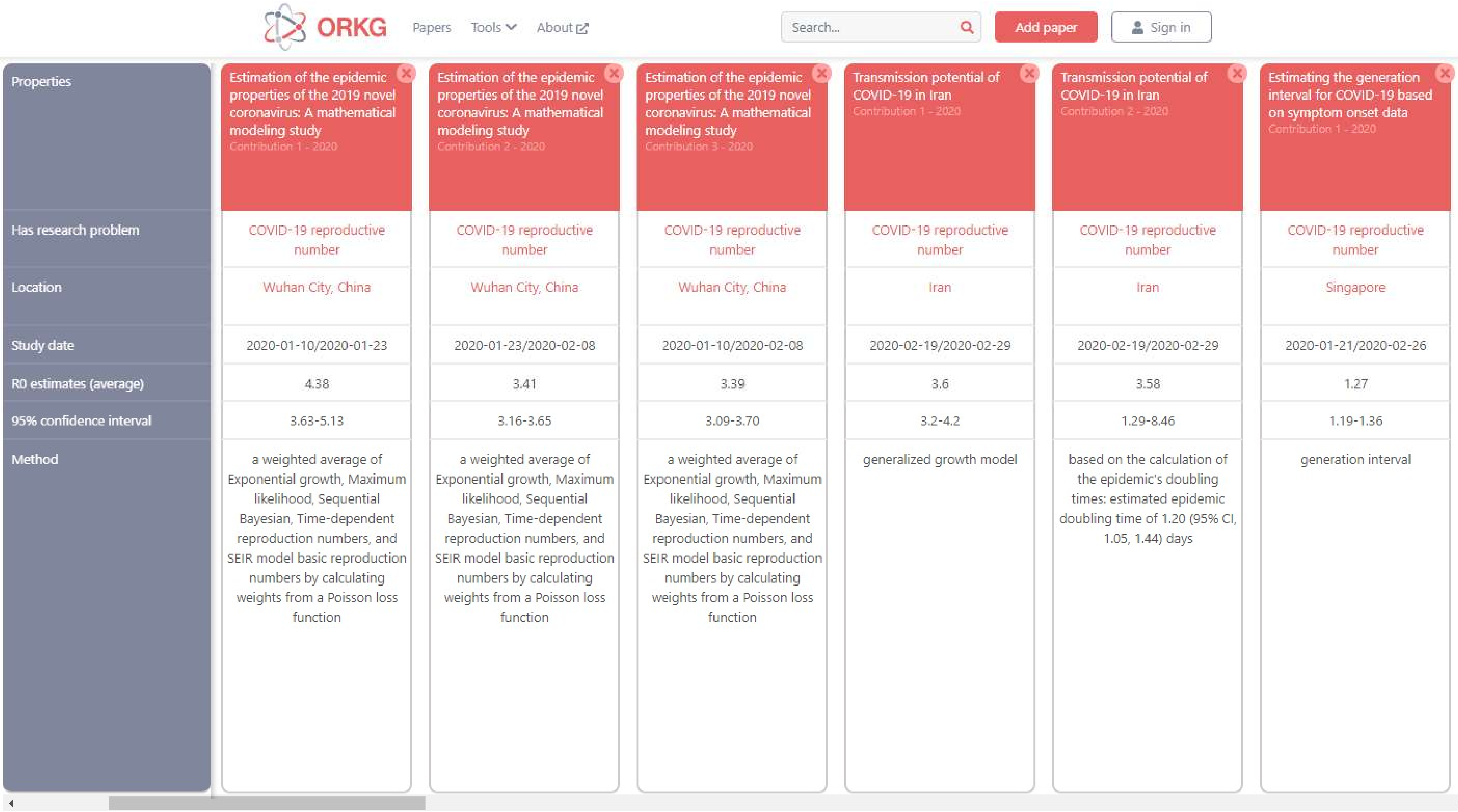
Q: How do different **genome editing** techniques **compare**?

Engineered Nucleases	Site-specificity	Safety	Ease-of-use / costs/ speed
zinc finger nucleases (ZFN)	++ 9-18nt	+	-- $$$: screening, testing to define efficiency
transcription activator-like effector nucleases (TALENs)	+++ 9-16nt	++	++ Easy to engineer 1 week / few hundred dollar
engineered meganucleases	+++ 12-40 nt	0	-- $$$ Protein engineering, high-throughput screening
CRISPR system/cas9	++ 5-12 nt	-	+++ Easy to engineer few days / less 200 dollar

LEIBNIZ INFORMATION CENTRE
FOR SCIENCE AND TECHNOLOGY
UNIVERSITY LIBRARY

TIB

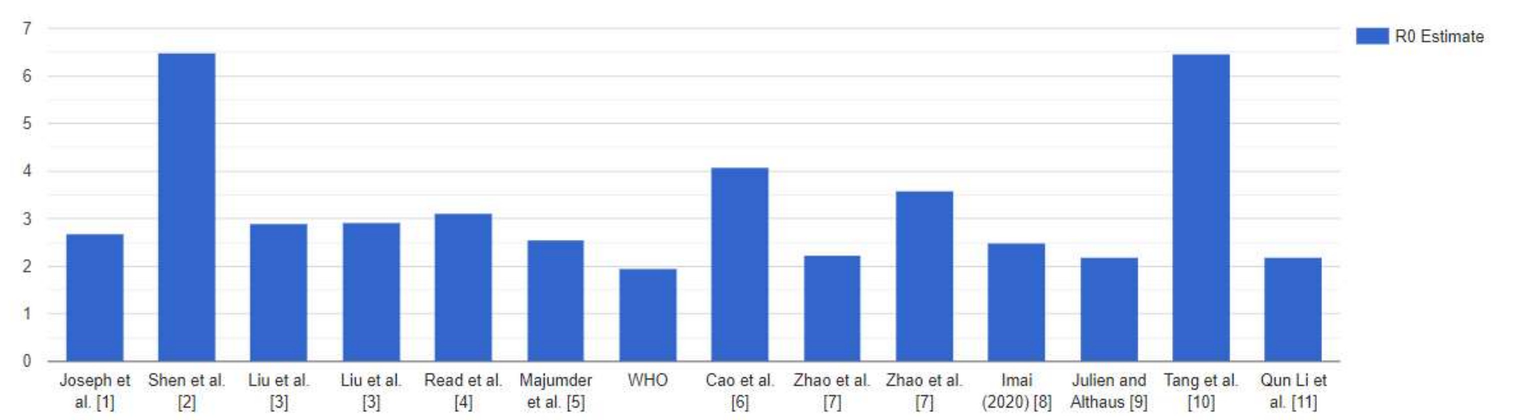
Demo: Open Research Knowledge Graph Prototype



ORKG Papers Tools About Search... Add paper Sign in

Properties	Estimation of the epidemic properties of the 2019 novel coronavirus: A mathematical modeling study Contribution 1 - 2020	Estimation of the epidemic properties of the 2019 novel coronavirus: A mathematical modeling study Contribution 2 - 2020	Estimation of the epidemic properties of the 2019 novel coronavirus: A mathematical modeling study Contribution 3 - 2020	Transmission potential of COVID-19 in Iran Contribution 1 - 2020	Transmission potential of COVID-19 in Iran Contribution 2 - 2020	Estimating the generation interval for COVID-19 based on symptom onset data Contribution 1 - 2020
Has research problem	COVID-19 reproductive number	COVID-19 reproductive number	COVID-19 reproductive number	COVID-19 reproductive number	COVID-19 reproductive number	COVID-19 reproductive number
Location	Wuhan City, China	Wuhan City, China	Wuhan City, China	Iran	Iran	Singapore
Study date	2020-01-10/2020-01-23	2020-01-23/2020-02-08	2020-01-10/2020-02-08	2020-02-19/2020-02-29	2020-02-19/2020-02-29	2020-01-21/2020-02-26
R0 estimates (average)	4.38	3.41	3.39	3.6	3.58	1.27
95% confidence interval	3.63-5.13	3.16-3.65	3.09-3.70	3.2-4.2	1.29-8.46	1.19-1.36
Method	a weighted average of Exponential growth, Maximum likelihood, Sequential Bayesian, Time-dependent reproduction numbers, and SEIR model basic reproduction numbers by calculating weights from a Poisson loss function	a weighted average of Exponential growth, Maximum likelihood, Sequential Bayesian, Time-dependent reproduction numbers, and SEIR model basic reproduction numbers by calculating weights from a Poisson loss function	a weighted average of Exponential growth, Maximum likelihood, Sequential Bayesian, Time-dependent reproduction numbers, and SEIR model basic reproduction numbers by calculating weights from a Poisson loss function	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

Bar Chart



Combo Chart

Conclusions

- We need to **reinvent scholarly communication**
- **Knowledge Graphs are perfectly suited to capture research contributions in a structured and semantic way making them human and machine interpretable**
- With our **Open Research Knowledge Graph initiative we aim to establish a registry for research contributions** (maybe similar to some extend as Crossref for DOIs/bibliographic metadata)
- Curation and **synergistic combination of human, expert and machine intelligence** is a challenge

Stay tuned

- https://tib.eu
- Mailinglist/group: https://groups.google.com/forum/#!forum/orkg
- Open Research Knowledge Graph: https://orkg.org
- ERC Consolidator Grant ScienceGRAPH started in May

LEIBNIZ INFORMATION CENTRE
FOR SCIENCE AND TECHNOLOGY
UNIVERSITY LIBRARY

TIB

https://de.linkedin.com/in/soerenauer

https://twitter.com/soerenauer

https://www.xing.com/profile/Soeren_Auer

http://www.researchgate.net/profile/Soeren_Auer

Contact
Prof. Dr. Sören Auer
TIB & Leibniz University of Hannover
auer@tib.eu

Leibniz Universität Hannover

Leibniz Association