Organizing Scientific Contributions with the Open Research Knowledge Graph

Prof. Dr. Sören Auer
Leibniz University of Hannover
TIB Leibniz Information Centre for Science and Technology
How did information flows change in the digital era?
How does it work today?

The World of Publishing & Communication has profoundly 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
What about Scholarly Communication?
Scholarly Communication has not changed (much)

17th century

19th century

20th century

21st century

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

Barbara Coscia1,2, Anil-Cyrielle Champagne3, Michael Rüber1, Daniel Grier1, Sandro Altobello Coelho, Sören Auer1, and Andreas Böth1

1 University of Leipzig, Germany  ·  2 B.D. & Ueber GmbH, Germany  ·  3 Federal University of João de Faria, Brazil  ·  4 University of Bonn & Fondazione IMS, Germany
email: {barbara,georg}@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 not only requires scalable and accurate approaches for the extraction of structured data in RDF (Resource Description Framework) from those webs. 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-augmented approach for named entity disambiguation. Our approach combines the Entity-indexed Topic Search (ETS) algorithm with label expansion strategies and string similarity measures. Based on this combination, AGDISTIS can efficiently detect the correct URI 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 surpass the state-of-the-art approach by up to 20% F-measure.

1 Introduction

The vision behind the Web of Data is to provide a new machine-readable layer on the Web where the content of Web pages is annotated with structured data (e.g., RDFa [3]). 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. Re-creating the vision of a machine-readable input layer of data that makes the Web more machine-readable is not only desirable, but also realistic: The potential of web-based annotation and retrieval is enormous. To this end, we need to annotate web pages with structured data, which can then be interlinked through the use of semantic technologies. This way, the Semantic Web (SW) can be realized, which promises to make the Web more accessible to people, machines, and organizations.

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:

<table>
<thead>
<tr>
<th>Digitalisation of Science</th>
<th>Monopolisation by commercial actors</th>
<th>Reproducibility Crisis</th>
<th>Proliferation of publications</th>
<th>Deficiency of Peer Review</th>
</tr>
</thead>
<tbody>
<tr>
<td>▪ Data integration and analysis</td>
<td>▪ Publisher look-in effects</td>
<td>▪ Majority of experiments are hard or not reproducible</td>
<td>▪ Publication output doubled within a decade</td>
<td>▪ Deteriorating quality</td>
</tr>
</tbody>
</table>

Duplication and Inefficiency

How can we avoid duplication if the terminology, research problems, approaches, methods, characteristics, evaluations, … are not properly defined and identified?

How would you build an engine / building without properly defining their parts, relationships, materials, characteristics …?
Root Cause –
Deficiency of Scholarly Communication?

Lack of…

<table>
<thead>
<tr>
<th>Transparency</th>
<th>Integratability</th>
<th>Machine assistance</th>
</tr>
</thead>
<tbody>
<tr>
<td>information is hidden in text</td>
<td>fitting different research results together</td>
<td>unstructured content is hard to process</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Identifyability</th>
<th>Collaboration</th>
<th>Overview</th>
</tr>
</thead>
<tbody>
<tr>
<td>of concepts beyond metadata</td>
<td>one brain barrier</td>
<td>Scientists look for the needle in the haystack</td>
</tr>
</tbody>
</table>
How good is CRISPR (wrt. precision, safety, cost)?

What specifics has genome editing with insects?

Who has applied it to butterflies?
How can we fix it?
Concepts

**Overarching Concepts**
- Research problems
- Definitions
- Research approaches
- Methods

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

**Domain specific Concepts**

<table>
<thead>
<tr>
<th>Mathematics</th>
<th>Physics</th>
<th>Chemistry</th>
<th>Computer Science</th>
<th>Technology</th>
<th>Architecture</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Definitions</td>
<td>- Experiments</td>
<td>- Substances</td>
<td>- Concepts</td>
<td>- Standards</td>
<td>- Regulations</td>
</tr>
<tr>
<td>- Theorems</td>
<td>- Data</td>
<td>- Structures</td>
<td>- Implementations</td>
<td>- Processes</td>
<td>- Elements</td>
</tr>
<tr>
<td>- Proofs</td>
<td>- Models</td>
<td>- Reactions</td>
<td>- Evaluations</td>
<td>- Elements</td>
<td>- Models</td>
</tr>
<tr>
<td>- Methods</td>
<td>- ...</td>
<td>- ...</td>
<td>- ...</td>
<td>- Units, Sensor data</td>
<td>- ...</td>
</tr>
<tr>
<td>- ...</td>
<td>- ...</td>
<td>- ...</td>
<td>- ...</td>
<td>- ...</td>
<td>- ...</td>
</tr>
</tbody>
</table>
Knowledge Graphs built on FAIR and Linked Data Principles

1. Use **URIs** to identify the “things” in your data

2. Use **http:// URIs** so people (and machines) can **look** them up on the web

3. When a URI is looked up, **return** a description of the thing in the W3C Resource Description Format (RDF)

4. Include **links to related things**

http://www.w3.org/DesignIssues/LinkedData.html
Chemistry Example: CRISPR Genome Editing

A practical guide to CRISPR/Cas9 genome editing in Lepidoptera

Linnin Zhang, Robert Reed
doi: https://doi.org/10.1101/130344

Now published in Diversity and Evolution of Butterfly Wing Patterns doi: 10.1007/978-981-10-4956-6_8

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 insertions in several lepidopteran species.
Chemistry Example: Populating the Graph

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

CRISPR / cas9 editing in Lepidoptera
https://doi.org/10.1101/130344

Robert Reed
https://orcid.org/0000-0002-6065-6728

addresses

CRSPRS/cas9

Genome editing in Lepidoptera

CRISPR/cas9 isEvaluatedWith

isImplementedBy

subProblemOf

isAuthorOf

Experimental Data

https://doi.org/10.5281/zenodo.896916

Genome editing

https://www.wikidata.org/wiki/Q24630389
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?

Result:
Automatic Generation of Comparisons / Surveys

Q: How do different genome editing techniques compare?

<table>
<thead>
<tr>
<th>Engineered Nucleases</th>
<th>Site-specificity</th>
<th>Safety</th>
<th>Ease-of-use / costs/ speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>zinc finger nucleases (ZFN)</td>
<td>++ 9-18nt</td>
<td>+</td>
<td>$$$: screening, testing to define efficiency</td>
</tr>
<tr>
<td>transcription activator-like effector nucleases (TALENs)</td>
<td>+++ 9-16nt</td>
<td>++</td>
<td>++ Easy to engineer 1 week / few hundred dollar</td>
</tr>
<tr>
<td>engineered meganucleases</td>
<td>+++ 12-40 nt</td>
<td>0</td>
<td>$$$ Protein engineering, high-throughput screening</td>
</tr>
<tr>
<td>CRISPR system/cas9</td>
<td>++ 5-12 nt</td>
<td>-</td>
<td>+++ Easy to engineer few days / less 200 dollar</td>
</tr>
</tbody>
</table>
Open Research Knowledge Graph
https://orkg.org
<table>
<thead>
<tr>
<th>Study</th>
<th>COVID-19 reproductive number</th>
<th>Location</th>
<th>Study date</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.38</td>
<td>Wuhan City, China</td>
<td>2020-01-10/2020-01-23</td>
<td>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</td>
</tr>
<tr>
<td>2</td>
<td>3.41</td>
<td>Wuhan City, China</td>
<td>2020-01-23/2020-02-08</td>
<td>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</td>
</tr>
<tr>
<td>3</td>
<td>3.39</td>
<td>Wuhan City, China</td>
<td>2020-01-10/2020-02-08</td>
<td>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</td>
</tr>
<tr>
<td>4</td>
<td>3.09</td>
<td>Iran</td>
<td>2020-02-19/2020-02-29</td>
<td>generalized growth model</td>
</tr>
<tr>
<td>5</td>
<td>3.6</td>
<td>Iran</td>
<td>2020-02-19/2020-02-29</td>
<td>based on the calculation of the epidemic doubling times: estimated epidemic doubling time of 1.20 (95% CI, 1.05, 1.44) days</td>
</tr>
<tr>
<td>6</td>
<td>3.58</td>
<td>Iran</td>
<td>2020-02-19/2020-02-29</td>
<td>1.27</td>
</tr>
<tr>
<td>7</td>
<td>1.29</td>
<td>Singapore</td>
<td>2020-01-21/2020-02-25</td>
<td>1.19-1.36</td>
</tr>
</tbody>
</table>
Transmission interval estimates suggest pre-symptomatic spread of COVID-19

DOI: 10.1101/2020.03.03.20023903

<table>
<thead>
<tr>
<th>Contribution 1</th>
<th>Contribution 2</th>
</tr>
</thead>
</table>

**Research problems**

COVID-19 reproductive number

**Contribution data**

<table>
<thead>
<tr>
<th>95% Confidence interval</th>
<th>1.45-2.48</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location</td>
<td>Singapore</td>
</tr>
<tr>
<td>Mean incubation period</td>
<td>7.1 (6.13, 8.25) days</td>
</tr>
<tr>
<td>Mean serial interval</td>
<td>4.56 (2.69, 6.42) days</td>
</tr>
<tr>
<td>R0 estimates (average)</td>
<td>1.97</td>
</tr>
<tr>
<td>Study date</td>
<td>2020-01-19/2020-02-26</td>
</tr>
</tbody>
</table>
Research problems
- COVID-19 reproductive number

Contribution data
- 95% Confidence interval: 1.45-2.48
- Location: Singapore
- Mean incubation period: 7.1 (6.13, 8.25) days
- Mean serial interval: 4.56 (2.69, 6.42) days
Research field overviews organize research problems, comparisons, visualizations etc.

Show latest updates, contributors, observatories
Domain Observatories

Make organizations supporting the curation of research in a certain domain more visible
Observatories are curated by organizations and organize research problems. Contributions, comparisons and visualizations in a particular field.
Showcase contributing organizations
The Team

Group Leaders

Dr. Markus Stocker
Dr. Gábor Kismihók
Dr. Javad Chamanara
Dr. Jennifer D'Souza

PostDocs

Dr. Lars Vogt
Dr. Oliver Karras

Doctoral Researchers

Allard Oelen
Yaser Jaradeh
Vitalis Wiens
Muhammad Haris

Software Development

Manuel Prinz
Kheir Eddine Farfar
Suhas Murthy

Collaborators InfAI Leipzig / AKSW

Golsa Heidary
Salomon Kabongo
Dr. Michael Martin
Natanael Arndt

Administration

Katja Bartel
Simone Matern
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
- Curation and **synergistic combination of human, expert and machine intelligence** is a challenge

Stay tuned

- [https://tib.eu](https://tib.eu)
- Consider creating an ORKG observatory for your domain
- Mailinglist/group: [https://groups.google.com/forum/#!forum/orkg](https://groups.google.com/forum/#!forum/orkg)
- Open Research Knowledge Graph: [https://orkg.org](https://orkg.org)
- ERC Consolidator Grant ScienceGRAPH on the topic
Contact
Prof. Dr. Sören Auer
TIB & Leibniz University of Hannover
auer@tib.eu

https://de.linkedin.com/in/soerenauer
https://twitter.com/soerenauer
https://www.xing.com/profile/Soeren_Auer
http://www.researchgate.net/profile/Soeren_Auer