Bioschemas

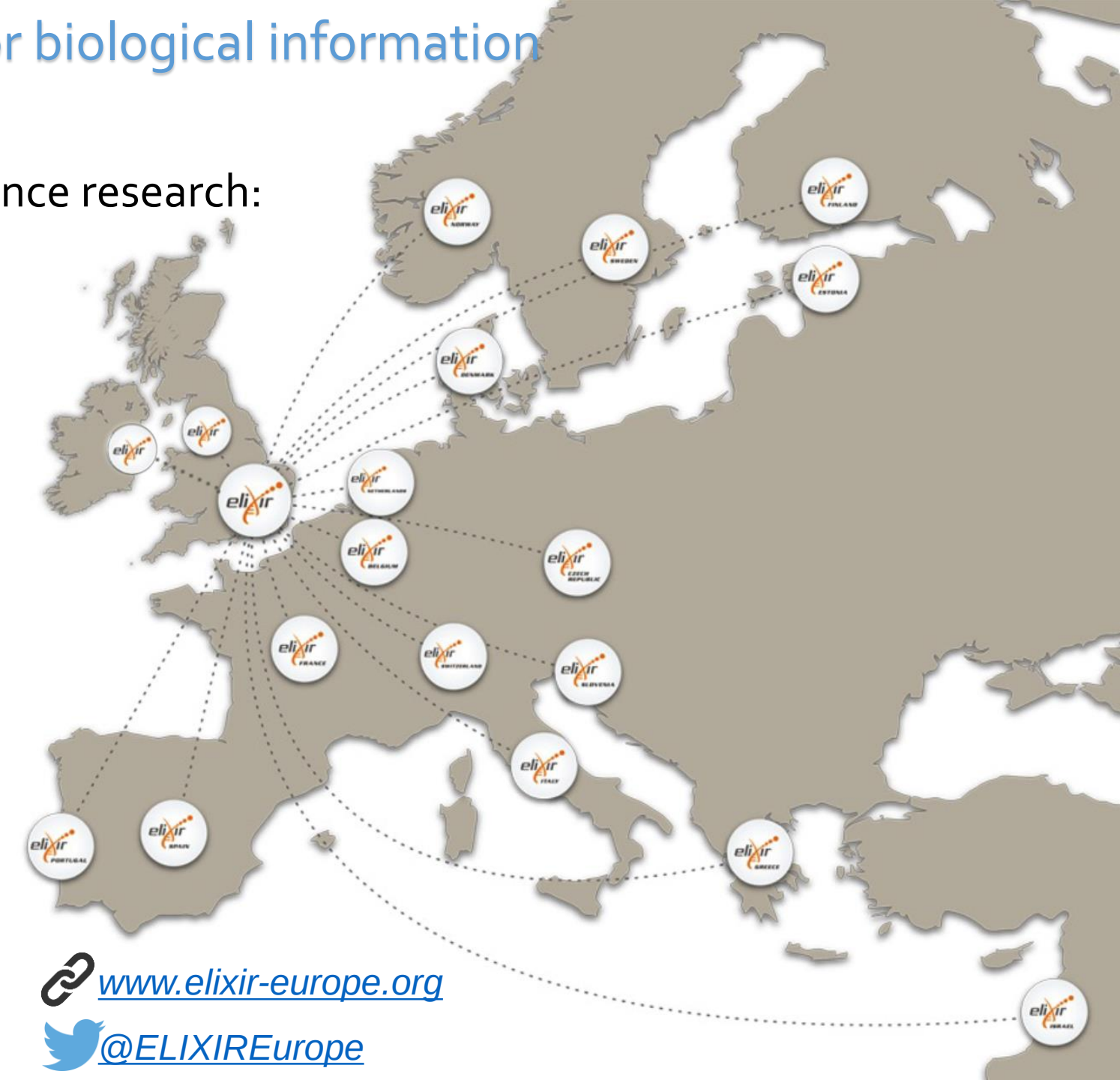
Structured data markup for life sciences with Schema.org

Roberto Preste

ELIXIR: European infrastructure for biological information

Data infrastructure for Europe's life-science research:

- *Data*
- *Interoperability*
- *Tools*
- *Compute*
- *Training*
- *Marine metagenomics*
- *Crop and forest plants*
- *Human data*
- *Rare diseases*



www.elixir-europe.org

@ELIXIREurope

Bioschemas

Research and information sharing

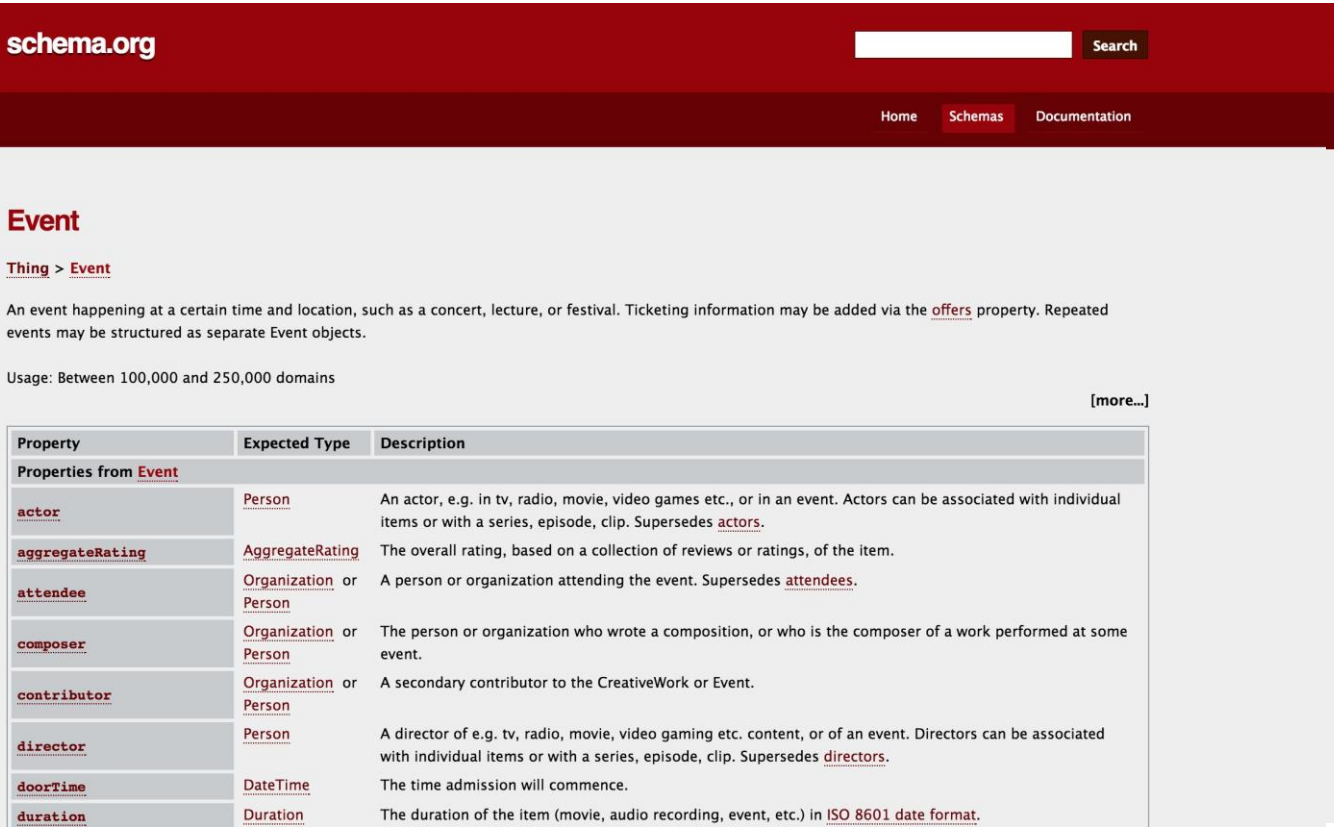
Information is exposed in the Internet through web pages

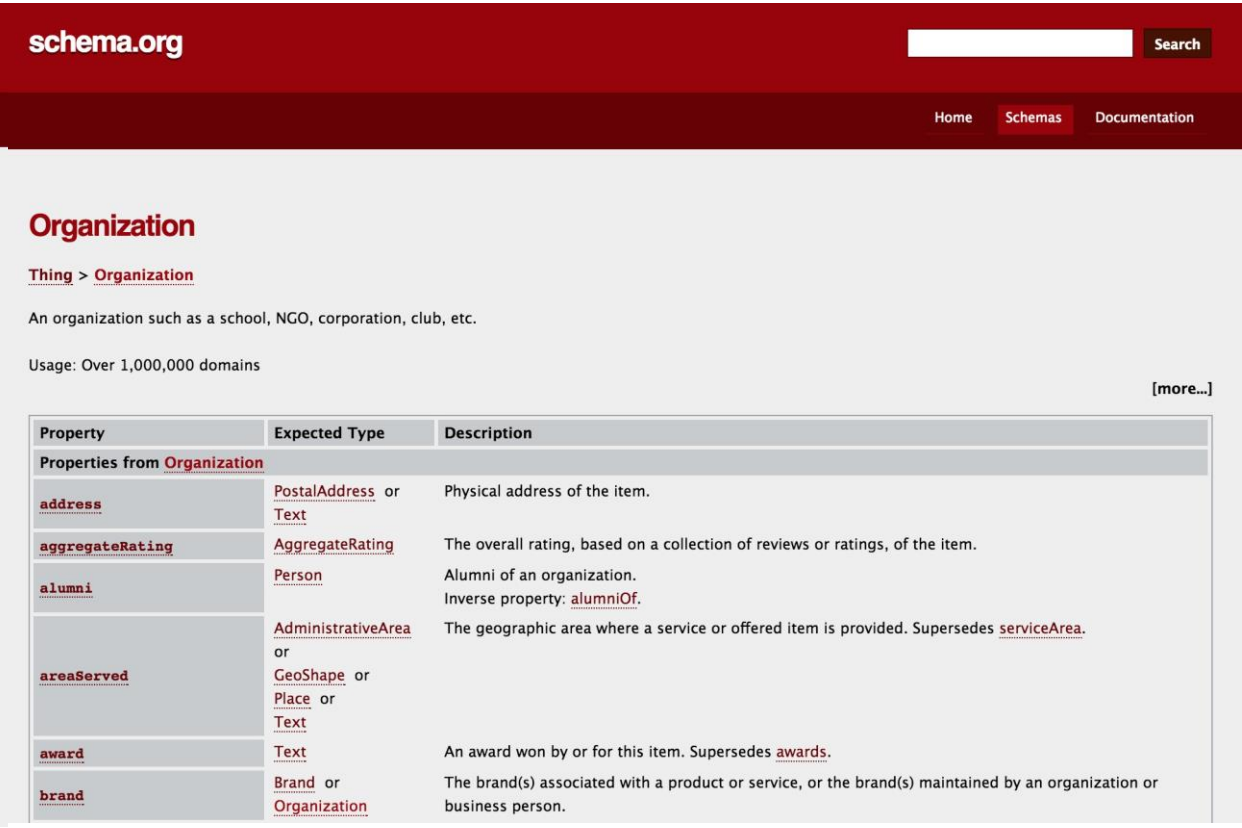
Great for common users

Bad for automated and programmatical data collection

Bioschemas

Schema.org



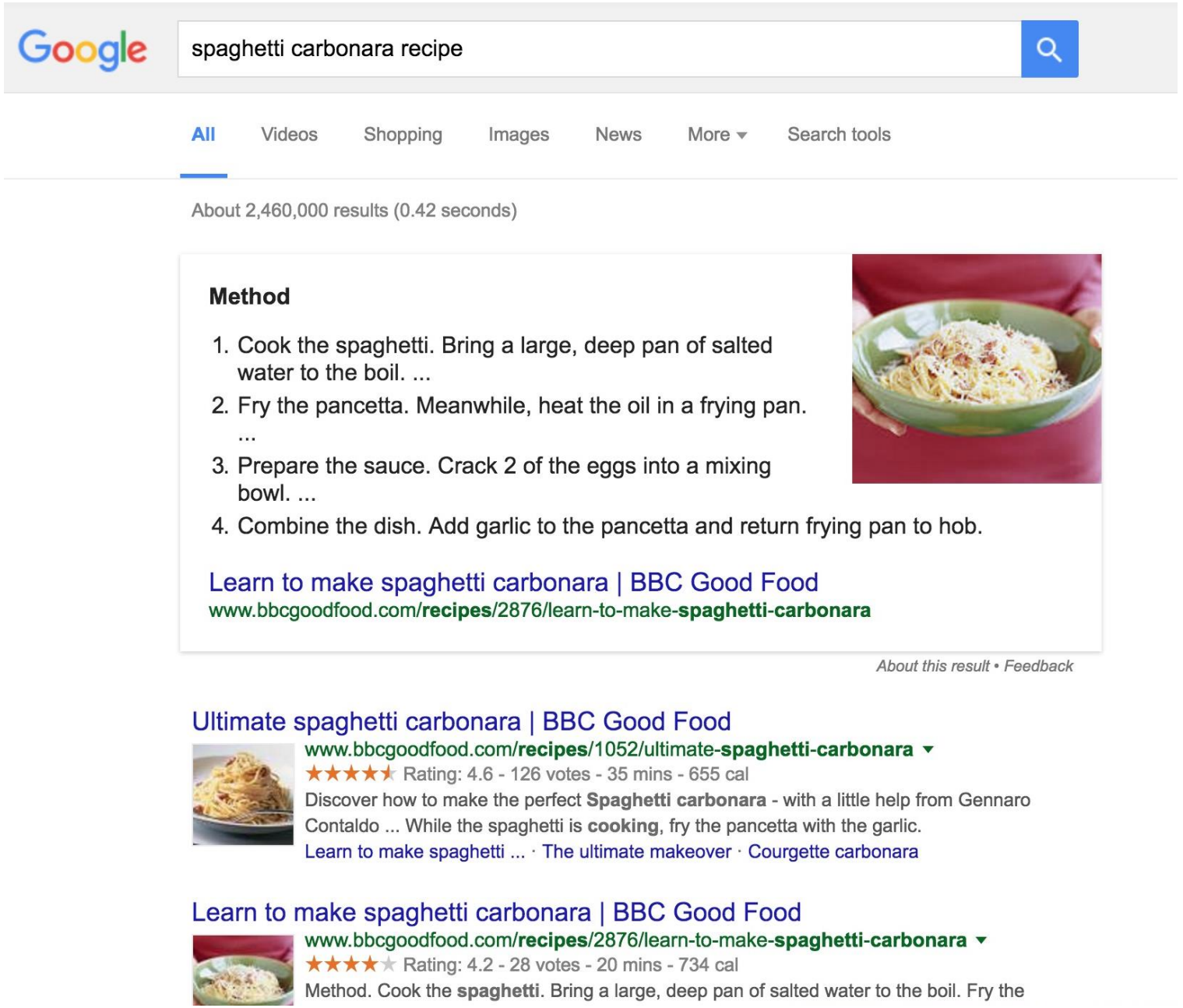


Community and collaborative initiative to integrate structured data markup on web pages, email messages and on the Internet in general, without altering the presentation layer

Bioschemas

Schema.org

Allows search engines to easily access websites' content and provide more useful results based on their underlying metadata



Bioschemas

Schema.org

Adopted by > 10 million websites

Sponsored by

...but still not this popular in the life science community!

Bioschemas

Bioschemas

Bioschemas aims to apply the Schema.org markup in life science

Better discoverability and usability of structured data

Bioschemas

Bioschemas: organization

Open community initiative:

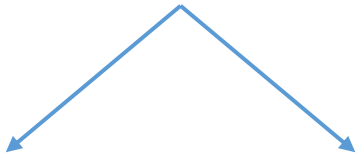
Organized in groups:

Group	Email	Email archive	Tasks/Issues	Shared Folder
Event	events@bioschemas.org			
Training Material	training-material@bioschemas.org			
Organization	organization@bioschemas.org			
Person	person@bioschemas.org			
Standard	standard@bioschemas.org			
Data	data@bioschemas.org			
Tool	tool@bioschemas.org			
Data Repositories	datarepositories@bioschemas.org			
Community	community@bioschemas.org			
Commons	commons@bioschemas.org			

Bioschemas

Bioschemas: specifications

Content types



Generic content types

E.g.: training materials, datasets

Can be used in many disciplines, besides life science

Biological content types

E.g.: pathways, proteins

Are domain-specific

In collaboration with health-lifesci.schema.org to integrate and extend

Schema.org types and properties

Bioschemas

Bioschemas: specifications

- ***Introduction***
- Data model
- Content guidelines
- Cardinality and controlled vocabularies
- New properties
- Examples

Overview of the problem, goal of the specification, technologies used to implement the solution

Problem statement

Conferences, workshops, meetings and events in general play a very important role in knowledge sharing and acquisition of new skills. Though there are several technological solutions for event sharing, events in life sciences are still not described in a consistent manner that would make them easy to discover, exchange or compare.

Because there is no single standard to follow, the dissemination, discovery and aggregation of events is not effective. As a result, the advertisement of an event in third party websites normally requires manual curation to shape the content to each provider's requirements.

Proposed solution

Rationale

In the development of the Life Science Events standard we have considered the following design goals.

Consensus

Many organisations and repositories providing events already exist. It is important this standard takes into account their experience and contribution.

Adoption

Many organisations already have a website or system providing information about events. They will not be willing to change their methods unless there is a clear benefit and a low barrier for adoption.

Reuse

There are existing formats and technologies suitable to represent at least some information about events. This specification will avoid reinvention and seek to extend existing standards.

Bioschemas

Bioschemas: specifications

- Introduction
- ***Data model***
- Content guidelines
- Cardinality and controlled vocabularies
- New properties
- Examples

Describes how to elaborate and implement more specific guidelines into a Schema.org skeleton

Data model

The data model proposed involves:

1. **Adopting the Schema.org Event type, and extending it with additional properties**. Schema.org already has a way of describing events, through its Event type. In this document we suggest using this type to describe life science events, but we also suggest new properties for this type, so that event descriptions can be more accurate and useful in life sciences. If the community agrees, these additional properties will be put forward for adoption to Schema.org.
2. **Adopting a standard way of using the Schema.org Event type.** Many properties in the Schema.org Event type are loosely defined, and we propose guidelines on how to use them so that they are more specific and consistent. These guidelines include concepts not supported by Schema.org, such as cardinality, controlled vocabularies and content guidelines (minimum, optional and recommended fields). For example, we suggest the use of a controlled vocabulary based on the EDAM ontology for the 'topics' property. These recommendations will not be part of the Schema.org Event type, but are proposed as best practices in using that type in life science.

Bioschemas: specifications

- Introduction
- Data model
- ***Content guidelines***
- Cardinality and controlled vocabularies
- New properties
- Examples

Defines a set of Minimum fields that are mandatory, Recommended fields for optimal discovery and integration, and Optional fields to enhance the user experience

Bioschemas

Event type definition

Data fields

Legend:

CN: Cardinality (one, many)
CG: Content Guideline (M: minimum; R: recommended; O: optional)
CV: Controlled Vocabulary

Property	Expected Type	Description	CN	CG	CV
Existing properties in schema.org/Event					
aggregateRating	AggregateRating	The overall rating, based on a collection of reviews or ratings, of the item.			
attendee	Organization or Person	A person or organization attending the event. Supersedes attendees.			
doorTime	DateTime	The time admission will commence.			
duration	Duration	The duration of the item (movie, audio recording, event, etc.) in ISO 8601 date format.			
endDate	Date	The end date and time of the item (in ISO 8601 date format).	One	M	
eventStatus	EventStatusType	An eventStatus of an event represents its status; particularly useful when an event is cancelled or rescheduled.			x
offers	Offer	An offer to provide this item—for example, an offer to sell a product, rent the DVD of a movie, or give away tickets to an event. Can use eligibleCustomerType and eligibleDuration properties to express any special offers.	One	O	
organizer	Organization or Person	An organizer of an Event.	Many	R	
performer	Organization or Person	A performer at the event—for example, a presenter, musician, musical group or actor. Supersedesperformers.	Many	O	

Bioschemas: specifications

- Introduction
- Data model
- Content guidelines
- ***Cardinality and controlled vocabularies***
- New properties
- Examples

Define the number of times a property can be instantiated and specific terms that need to be used

Event type definition

Data fields

Legend:

CN: Cardinality (one, many)
CG: Content Guideline (M: minimum; R: recommended; O: optional)
CV: Controlled Vocabulary

Property	Expected Type	Description	CN	CG	CV
Existing properties in schema.org/Event					
aggregateRating	AggregateRating	The overall rating, based on a collection of reviews or ratings, of the item.			
attendee	Organization or Person	A person or organization attending the event. Supersedes attendees.			
doorTime	DateTime	The time admission will commence.			
duration	Duration	The duration of the item (movie, audio recording, event, etc.) in ISO 8601 date format.			
endDate	Date	The end date and time of the item (in ISO 8601 date format).	One	M	
eventStatus	EventStatusType	An eventStatus of an event represents its status; particularly useful when an event is cancelled or rescheduled.			x
offers	Offer	An offer to provide this item—for example, an offer to sell a product, rent the DVD of a movie, or give away tickets to an event. Can use eligibleCustomerType and eligibleDuration properties to express any special offers.	One	O	
organizer	Organization or Person	An organizer of an Event.	Many	R	
performer	Organization or Person	A performer at the event—for example, a presenter, musician, musical group or actor. Supersedesperformers.	Many	O	

Bioschemas

Bioschemas: specifications

- Introduction
- Data model
- Content guidelines
- Cardinality and controlled vocabularies
- ***New properties***
- Examples

New properties can be suggested in order to provide more accurate and useful type description in life science

Proposed new properties for the Event type					
eventId	Text	Unique Id for the event.	One	O	
prerequisite	Text	A list of prerequisites to be able to attend the event.	Many	R	
accreditation	Organization or Text	Type of accreditation or organisation that accredits the event.	One	R	
eligibility	Text	Defines the type of eligibility to attend this event e.g first come first served.	Many	R	x
capacity	Integer	Available number of spaces	One	R	
contact	Organization or Person	Main point of contact that can be contacted for general queries. This would be an event organiser or an administrator.	Many	M	
attachment	URL	Any files or related websites which give more information about this event. e.g. flyers, third party sites handling tickets.	Many	O	

Bioschemas: specifications

- Introduction
- Data model
- Content guidelines
- Cardinality and controlled vocabularies
- New properties
- ***Examples***

No Markup | Microdata | RDFa | JSON-LD

```
...
Epigenomics of Common Diseases
This conference will bring together leading scientists from the fields of epigenomics, genetics and bioinformatics.
Event type: Workshops and courses
Date: Wednesday 15 April 2015
```

No Markup | **Microdata** | RDFa | JSON-LD

```
<div itemscope itemtype="http://schema.org/Event">
  ...
    <div itemprop="name">Epigenomics of Common Diseases</div>
    <div itemprop="description">This conference will bring together leading scier
    <div>Event type: <span itemprop="eventType">Workshops and courses</span></div
    <div>Date: <meta itemprop="startDate" content="2015-04-15T">Wednesday 15 Apri
  ...
```

No Markup | Microdata | **RDFa** | JSON-LD

```
<div vocab="http://schema.org/" typeof="Event">
  ...
    <div property="name">Epigenomics of Common Diseases</div>
    <div property="description">This conference will bring together leading scientists from the fields of epigenomic
    <div>Event type: <span property="eventType">Workshops and courses</span></div>
    <div>Date: <meta property="startDate" content="2015-04-15T">Wednesday 15 April 2015</div>
  ...
```

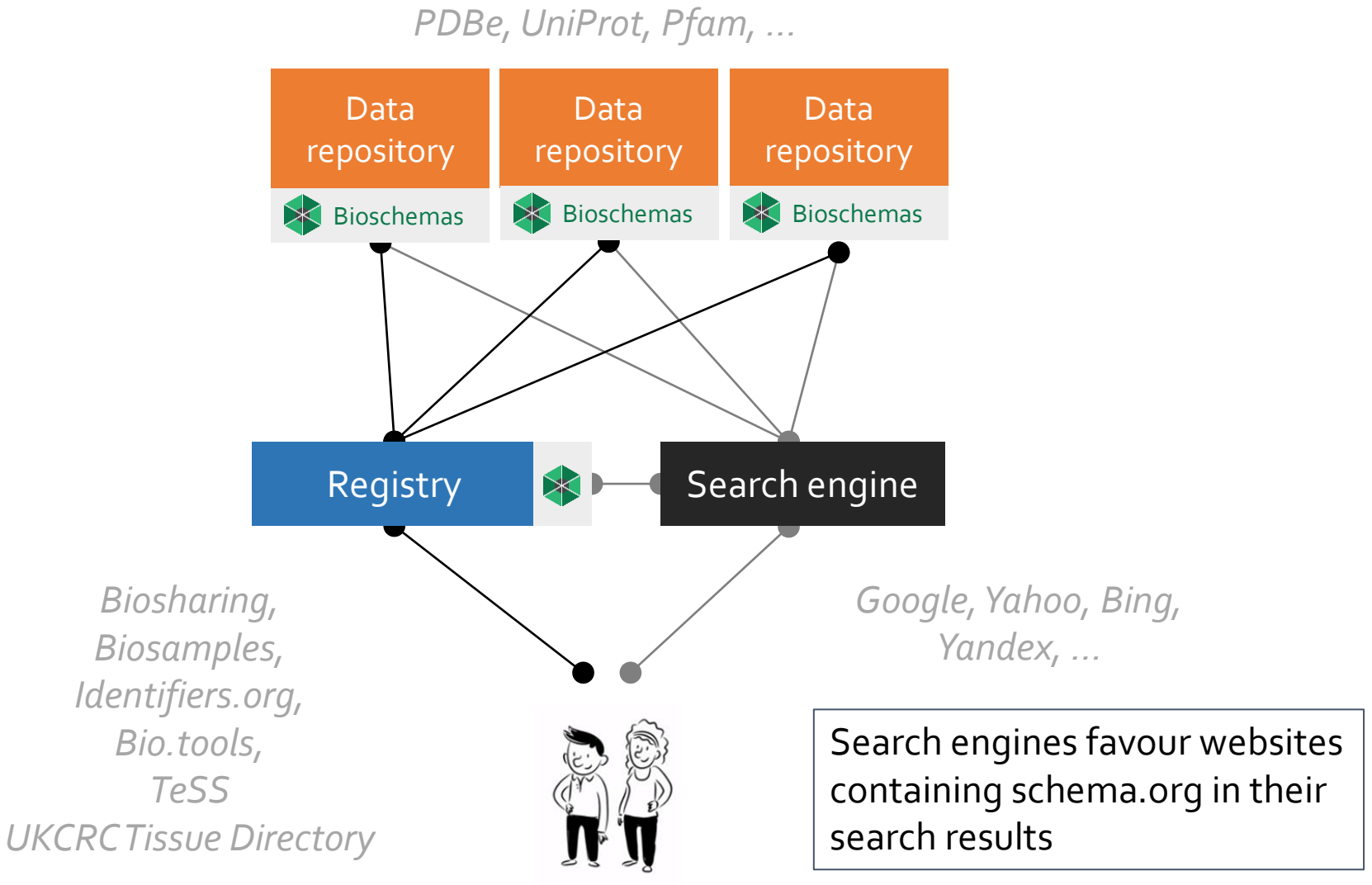
No Markup | Microdata | RDFa | **JSON-LD**

```
<script type="application/ld+json">
{
    "@context": "http://schema.org/",
    "@type": "Event",
    "name": "Epigenomics of Common Diseases",
    "description": "This conference will bring together leading scientists from the fields of epigenomics, genetics
    "eventType": "Workshops and courses",
```

Bioschemas

Bioschemas: applications

Improving search results via Bioschemas

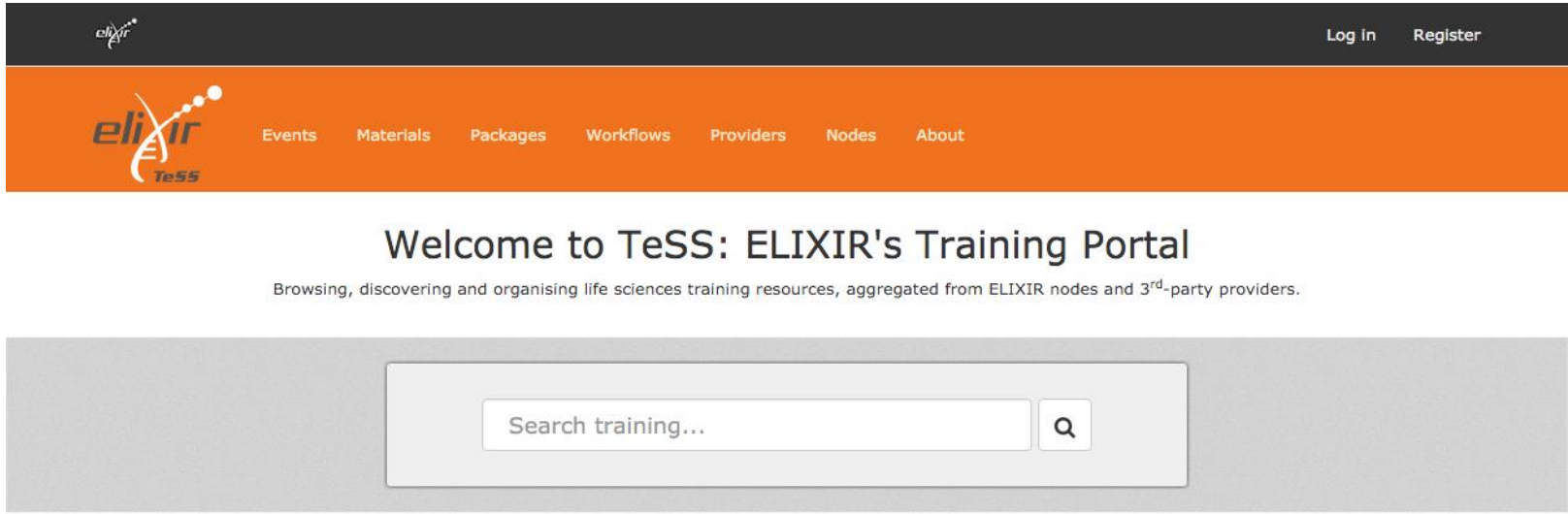


Bioschemas

Bioschemas: applications

TeSS: ELIXIR's Training Portal

https://tess.elixir-uk.org



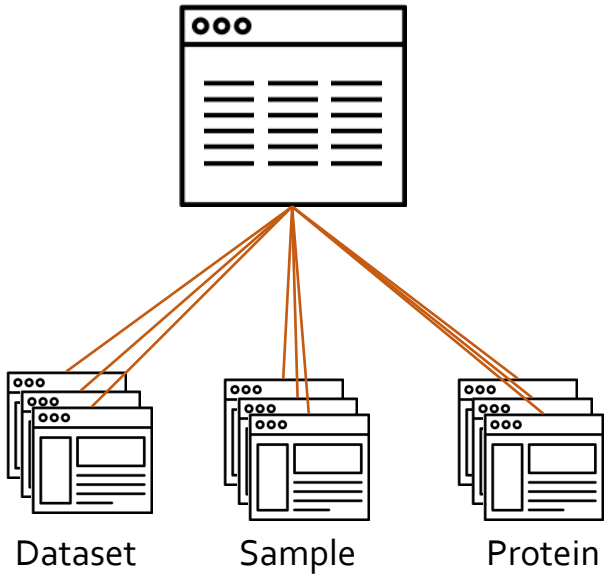
Bioschemas

Bioschemas: validation

Discovery, Registration and Validation

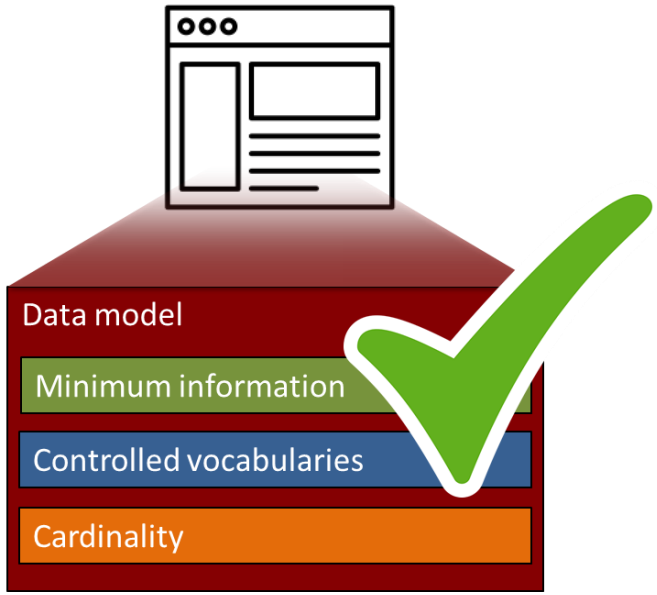
DISCOVERY

Provide a simple registry where providers or consumers can describe of sites providing content in schema.org compliant with Bioschemas



VALIDATION

Provide a GUI to validate Bioschemas compliant websites and Validate data repositories adopting Bioschemas



Bioschemas

Bioschemas: validation

Bioschemas
Registry & Validator

Home | Registry | Test Website | Submit Website

Test And Submit A Website

With the Submit tool you can test a specific URL for its compliance with the Bioschemas specifications and submit it to the Registry for public consultation. You can choose to test all the children links for that page (links with the same root URL) as well as scraping your URL for one specific Bioschemas type (Event, Organization, Person, Training Material).

Website URL:

Website Name:

If you choose not to provide a name for the URL you are testing, the *title* field content from the tested website will be used as its name

☐ Validate children links too

Scrape a specific Bioschemas type: All Types

☐ Mail notification when the process is done example@domain.com

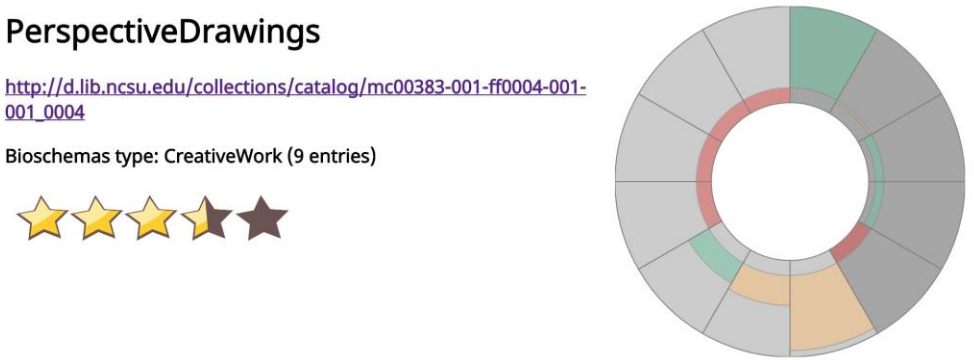
Validate!

Summary

PerspectiveDrawings

http://d.lib.ncsu.edu/collections/catalog/mc00383-001-ff0004-001-001_0004

Bioschemas type: CreativeWork (9 entries)



Minimum and Recommended Properties

☐ Show properties description

Compliance	Property	Type	Guideline	CV	Cardinality
93%	url	URL	Recommended	No	Many
100%	name	Text	Minimum	No	One
0%	license	URL or CreativeWork	Recommended	Yes	One
47%	keywords	Text	Recommended	No	Many
13%	genre	Text or URL	Minimum	Yes	Many
40%	description	Text	Recommended	No	One
7%	dateModified	Date or DateTime	Minimum	No	One
0%	contributor	Person or Organization	Recommended	No	Many
0%	author	Person or Organization	Recommended	No	Many
0%	audience	Audience	Minimum	Yes	Many
7%	about	Thing	Minimum	No	One
0%	level	Text	Recommended	Yes	One

Bioschemas

Upcoming activities

Implementation Study Objectives

Life sciences Content Types

Schema.org content types for life science Data
- Data repository, Dataset
- Samples, Protein annotations
- Phenotype annotations

Discovery and validation

- Publication of metadata
- Automated integration of metadata in specialised registries

Discovery and validation of Bioschemas entries

Community Support and Promotion

Support community and adoption
Alignment of technical activities
Working group within ELIXIR
Collaboration between ELIXIR and BD2K
Test benefits and issues

Hands-on workshops

Small delivery team

Bioschemas

Upcoming activities

- **Bioschemas AGM**, 8th-9th November, Rothamsted (UK)
  Details on https://goo.gl/hu7uYK

- More content types for life science in development:
  ~~Tools~~
  Data repository
  Dataset
  Sample
  Phenotype
  Protein annotations

Acknowledgements

Bioschemas community

Group chairs

- **Community**
  Premysl Velek
- **Event**
  Martin Cook
- **Training materials**
  Aleksandra Nenadic & Gabriella Rustici
- **Organization**
  Richard Holland & Rafael C Jimenez
- **Person**
  Niall Beard
- **Standard**
  A Gonzalez-Beltran & P McQuilton

Organization representatives

- **ELIXIR**
  Premysl Velek
- **Pistoia Alliance**
  Richard Holland
- **GOBLET**
  Terry Atwood
- **BBMRI**
  Michaela Mayrhofer
- **TeSS**
  Niall Beard
- **BioSharing**
  SA Sansone, A Gonzalez-Beltran, P McQuilton, P Rocca-Serra
- **NIH BD2K bioCADDIE**
  SA Sansone, A Gonzalez-Beltran, Jeff Grethe

Bioschemas

Contributors

- Aleksandra Nenadic
- Adam Hospital
- Gabriella Rustici
- Carlos Horro
- Martin Cook
- Niall Beard
- Rafael C Jimenez
- Andy Jenkinson
- Manuel Corpas
- Roberto Preste
- Richard Holland
- Alejandra Gonzalez-Beltran
- Andrew Lonie
- Carole Goble
- Peter McQuilton
- Premysl Velek
- Ian Dunlop
- Jef Grethe
- Milo Thurston
- Niklas Blomberg
- Yasset Perez-Riverol
- Sarala Wimalaratne
- Nick Juty
- Jose Luis Ambite
- Brane Leskošek
- Celia van Gelder
- Christa Janko
- Christine Staiger
- Dan Brickley
- Daniel Faria
- Dmitry Repchevsky
- Daniel Sobral
- Daniel Vaughan
- Ian Fore
- Frederik Coppens
- Josep Gelpi
- ChuQiao Gong
- Hedi Peterson
- Hervé Ménager
- Nina Hrtonova
- Isabelle Perseil
- Jaap Heringa
- Jon Ison
- John Hancock
- Simon Jupp
- John D. Van Horn
- Ivana Krenkova
- Laura Furlong
- Morris Swertz
- Mateusz Kuzak
- Mario Alberich
- Mark Thompson
- Maria Martin
- Mikael Borg
- Montserrat Gonzàlez
- Norman Morrison
- Nùria Queralt-Rosinach
- Olivier Sallou
- Robert Pergl
- Pedro Fernandes
- Pierre Larmande
- Rob Finn
- Renzo Kottmann
- Rodrigo Lopez
- Sameer Velankar
- Sara Light
- Carol Shreffler
- Silvano Squizzato
- Susanna Sansone
- Tony Burdett
- Terry Atwood
- Cath Brooksbank
- Luc Deltombe
- Michaela Mayrhofer
- Philippe Rocca-Serra

Bioschemas

Bioschemas

Thank you!

all@bioschemas.org

http://bioschemas.org

@BioSchemas