Bioschemas: From Potato Salad to Protein Annotation

Bioschemas:
From Potato Salad to Protein Annotation

Alasdair J G Gray[1,2], Carole Goble[1,3], Rafael C Jimenez[4], and
The Bioschemas Community[5]

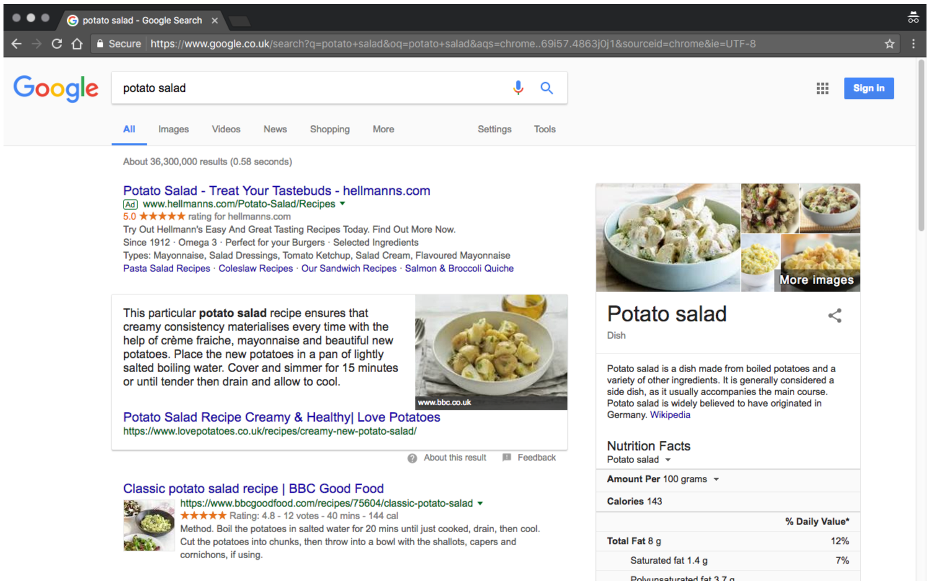
[1] ELIXIR-UK
[2] Heriot-Watt University, UK
[3] University of Manchester, UK
[4] ELIXIR-Hub, Hinxton Genome Campus, UK
[5] `http://bioschemas.org`

Abstract. The life sciences have a wealth of data resources with a wide range of overlapping content. Key repositories, such as UniProt for protein data or Entrez Gene for gene data, are well known and their content easily discovered through search engines. However, there is a long-tail of bespoke datasets with important content that are not so prominent in search results. Building on the success of Schema.org for making a wide range of structured web content more discoverable and interpretable, e.g. food recipes, the Bioschemas community (`http://bioschemas.org`) aim to make life sciences datasets more findable by encouraging data providers to embed Schema.org markup in their resources.

Keywords: Schema.org, metadata, dataset descriptions, data discovery

1 Introduction

Schema.org provides a way to add semantic markup to web pages to enable those web pages to become more interpretable by the search engines that index them, and therefore to improve search results [2]. Schema.org markup describes *types* of information, which then have *properties*. For example, `Recipe` is a type for representing cooking recipes that has properties like `cookTime`, `nutrition`, and `recipeIngredient` for marking up the characteristics of the recipe. Schema.org markup is increasingly being applied to web pages as it boosts a site's ranking in search results [2]. Schema.org markup in web pages also enhances the search experience for end users; enabling them to make more informed decisions when deciding between two search results. For example, when searching for a recipe for potato salad the result snippets contain information such as cooking time and the number of calories (see Fig. 1). These have been extracted from the Schema.org markup of the underlying web pages and enable the user to make a decision without reading the whole web page. Another example of services being built over Schema.org markup include the content of the knowledge graphs of the search engines (also shown in Fig. 1).


potato salad - Google Search
Secure
https://www.google.co.uk/search?q=potato+salad&oq=potato+salad&aqs=chrome..69i57.4863j0j1&sourceid=chrome&ie=UTF-8
Google
potato salad
Sign in
All
Images
Videos
News
Shopping
More
Settings
Tools
More images

About 36,300,000 results (0.58 seconds)

Potato Salad - Treat Your Tastebuds - hellmanns.com
Ad www.hellmanns.com/Potato-Salad/Recipes
5.0 ★★★★★ rating for hellmanns.com
Try Out Hellmann's Easy And Great Tasting Recipes Today. Find Out More Now.
Since 1912 · Omega 3 · Perfect for your Burgers · Selected Ingredients
Types: Mayonnaise, Salad Dressings, Tomato Ketchup, Salad Cream, Flavoured Mayonnaise
Pasta Salad Recipes · Coleslaw Recipes · Our Sandwich Recipes · Salmon & Broccoli Quiche

This particular **potato salad** recipe ensures that creamy consistency materialises every time with the help of crème fraiche, mayonnaise and beautiful new potatoes. Place the new potatoes in a pan of lightly salted boiling water. Cover and simmer for 15 minutes or until tender then drain and allow to cool.

www.bbc.co.uk

Potato Salad Recipe Creamy & Healthy| Love Potatoes
https://www.lovepotatoes.co.uk/recipes/creamy-new-potato-salad/

About this result Feedback

Classic potato salad recipe | BBC Good Food
https://www.bbcgoodfood.com/recipes/75604/classic-potato-salad
★★★★★ Rating: 4.8 - 12 votes - 40 mins - 144 cal
Method. Boil the potatoes in salted water for 20 mins until just cooked, drain, then cool. Cut the potatoes into chunks, then throw into a bowl with the shallots, capers and cornichons, if using.

Potato salad
Dish

Potato salad is a dish made from boiled potatoes and a variety of other ingredients. It is generally considered a side dish, as it usually accompanies the main course. Potato salad is widely believed to have originated in Germany. Wikipedia

Nutrition Facts
Potato salad

Amount Per 100 grams	
Calories 143	
	% Daily Value*
Total Fat 8 g	12%
Saturated fat 1.4 g	7%
Polyunsaturated fat 3.7 g	

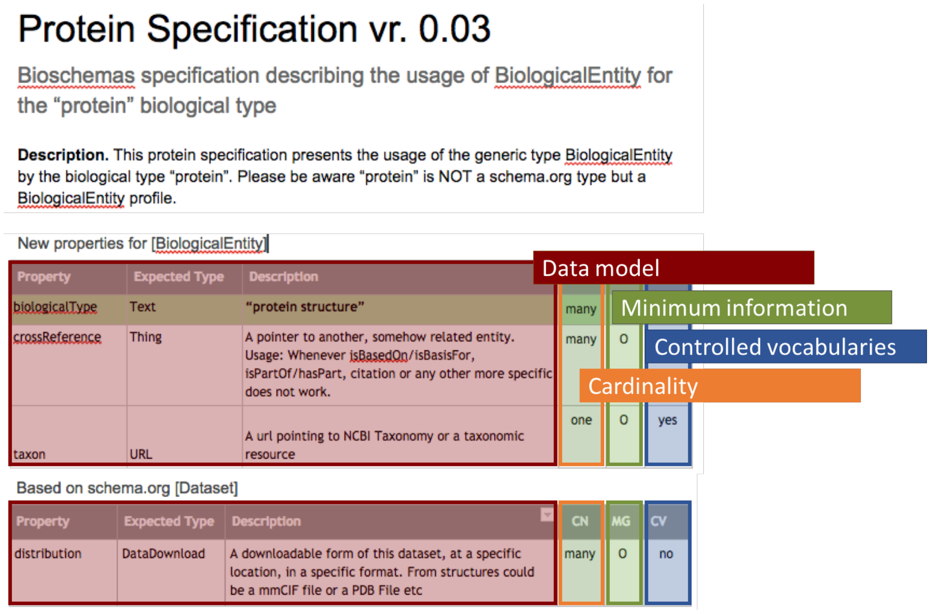
Fig. 1. Search result for potato salad showing rich snippets generated from Schema.org markup as well as content from the Google Knowledge Graph that has been populated with Schema.org markup.

The life sciences community have a wealth of data resources with a wide range of overlapping content. When gathering data about a particular gene or protein, scientists want the data to be aggregated from all available sources. Currently data from key repositories, such as UniProt (SIB/EBI) for information about proteins [3] or Gene (NIH) for information about genes [1], are well known and easily gathered. However, there is a long-tail of bespoke datasets with important content whose content are not so readily available. The Bioschemas community[6] aim to make life sciences datasets more findable by encouraging data providers to embed Schema.org markup in their resources. Thus enabling aggregation of content through a common approach that will enable novel applications. Previous work has added Biomedical terms[7], but these are not sufficient for the breadth of the life sciences community. The Bioschemas community are working in conjunction with the wider Schema.org community.

2 Bioschemas

Within the life sciences there is demand to discover more than just generic types like `Dataset` and `Event`. The Bioschemas community have identified a wide range of discovery use cases searching for different types of biological resources.

[6] `http://bioschemas.org` (accessed Sept 2017)
[7] `https://health-lifesci.schema.org/` (accessed Sept 2017)



Protein Specification vr. 0.03

Bioschemas specification describing the usage of BiologicalEntity for the "protein" biological type

Description. This protein specification presents the usage of the generic type BiologicalEntity by the biological type "protein". Please be aware "protein" is NOT a schema.org type but a BiologicalEntity profile.

New properties for [BiologicalEntity]

Property	Expected Type	Description			
biologicalType	Text	"protein structure"	many		
crossReference	Thing	A pointer to another, somehow related entity. Usage: Whenever isBasedOn/isBasisFor, isPartOf/hasPart, citation or any other more specific does not work.	many	O	
taxon	URL	A url pointing to NCBI Taxonomy or a taxonomic resource	one	O	yes

Based on schema.org [Dataset]

Property	Expected Type	Description	CN	MG	CV
distribution	DataDownload	A downloadable form of this dataset, at a specific location, in a specific format. From structures could be a mmCIF file or a PDB File etc	many	O	no

Fig. 2. The Protein Specification with the Bioschemas layers highlighting the Schema.org data model, expected cardinality of the property, minimum information recommendations, and controlled vocabularies.

These include searching for data about a specific biological entity such as a particular gene or protein, discovering a data repository to deposit experimental results, and identifying the storage location of specific biological samples[8]. Currently biological types like genes, proteins, and samples are not represented in Schema.org. Bioschemas aims to engage with life science communities relying on existing community agreements to bring forward new biological types to Schema.org.

For any given entity type in Schema.org there are a large number of properties available, many inherited from parent types. For example, `Dataset` has two properties (`distribution` and `includedInDataCatalog`) but inherits 78 properties from `CreativeWork` and 11 from `Thing`. This is far more properties than can be realistically expected from resource providers, c.f. [2]. Additionally, this wide range of choice makes it difficult to develop tools to consume markup. To support tools developed to exploit Schema.org markup in web resources, it is beneficial if the markup is done in a consistent way, i.e. all resources describing a particularly type of entity provide the same set of properties.

The Bioschema specifications are being developed in an example driven manner in a short timeframe – the ELIXIR Implementation Study runs for just one

[8] Links to documents containing these use cases can be found on the Bioschemas website `http://bioschemas.org/groups/` (accessed Sept 2017)

calendar year (2017). The Bioschema specifications go beyond simply extending Schema.org with new types and properties for biological entities. As shown in Fig. 2, the Bioschemas specifications layer provides additional constraints over the Schema.org model. These constraints capture (i) the minimal information properties agreed by the community which are mandatory (M), recommended (R), or optional (O), (ii) the cardinality of the property, i.e. whether it is expected to occur once or many times, and (iii) associated controlled vocabulary terms drawn from existing ontologies. Following from the experience of the wider Schema.org community [2], the Bioschemas specifications aim to require just 6 properties for any resource type. These properties are being selected based on their ability to support indexing and snippet generation to enable a consumer of the search result to discover and distinguish between resources.

3 Future Work

The Bioschemas Implementation Study is currently in the development/testing phase of its lifecycle. To ensure the viability of the specifications from the resource providers perspective, example deployments are being developed. At the same time, tools for consuming and exploiting the markup are also being developed. The outcome of both these development processes will feed into the final revisions of the specifications and proposed extensions to the core Schema.org vocabulary.

While the Bioschemas community has a primary focus on life sciences data, prominent members of the community are involved with the European Open Science Cloud project[9] with the aim to adopt the Bioschemas approach of defining community agreed Schema.org markup profiles in other scientific disciplines.

Acknowledgements

The work presented here represents the contributions of the whole Bioschemas Community (`http://bioschemas.org/people/`). The current work is funded through an ELIXIR Implementation Study (`https://www.elixir-europe.org/activities/bioschemas`) and the EU ELIXIR-EXCELERATE grant within the Research Infrastructures programme of Horizon 2020, grant agreement number 676559.

References

1. Brown et al, G.R.: Gene: a gene-centered information resource at ncbi. NAR 43(D1), D36–D42 (2015), `http://dx.doi.org/10.1093/nar/gku1055`
2. Guha, R.V., Brickley, D., Macbeth, S.: Big data makes common schemas even more necessary. CACM 59(2) (2016), `http://dx.doi.org/10.1145/2844544`
3. The UniProt Consortium: UniProt: the universal protein knowledgebase. NAR 45(D1), D158–D169 (2017), `http://dx.doi.org/10.1093/nar/gkw1099`

[9] European Open Science Cloud `https://eoscpilot.eu/` accessed Sept 2017