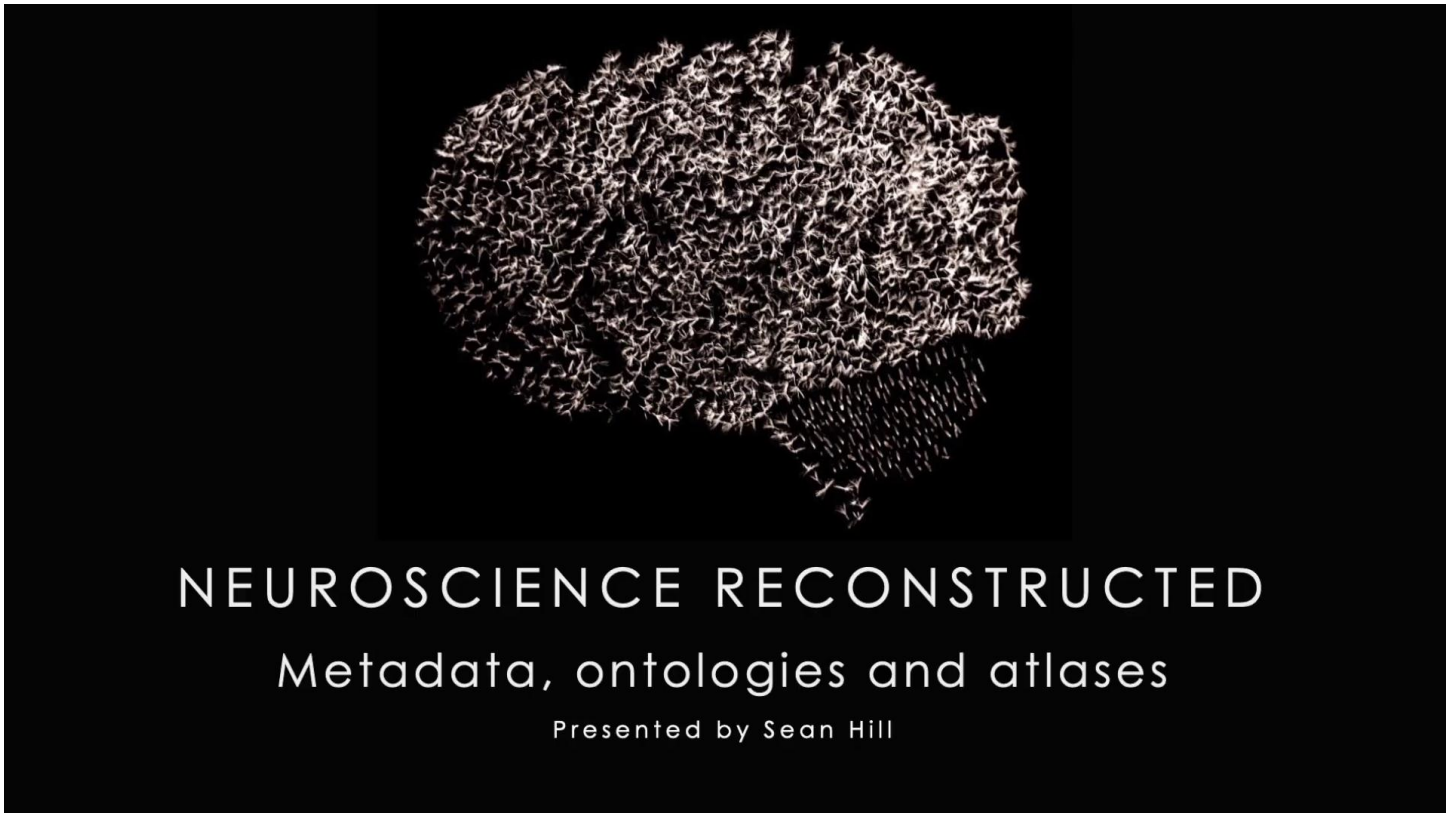
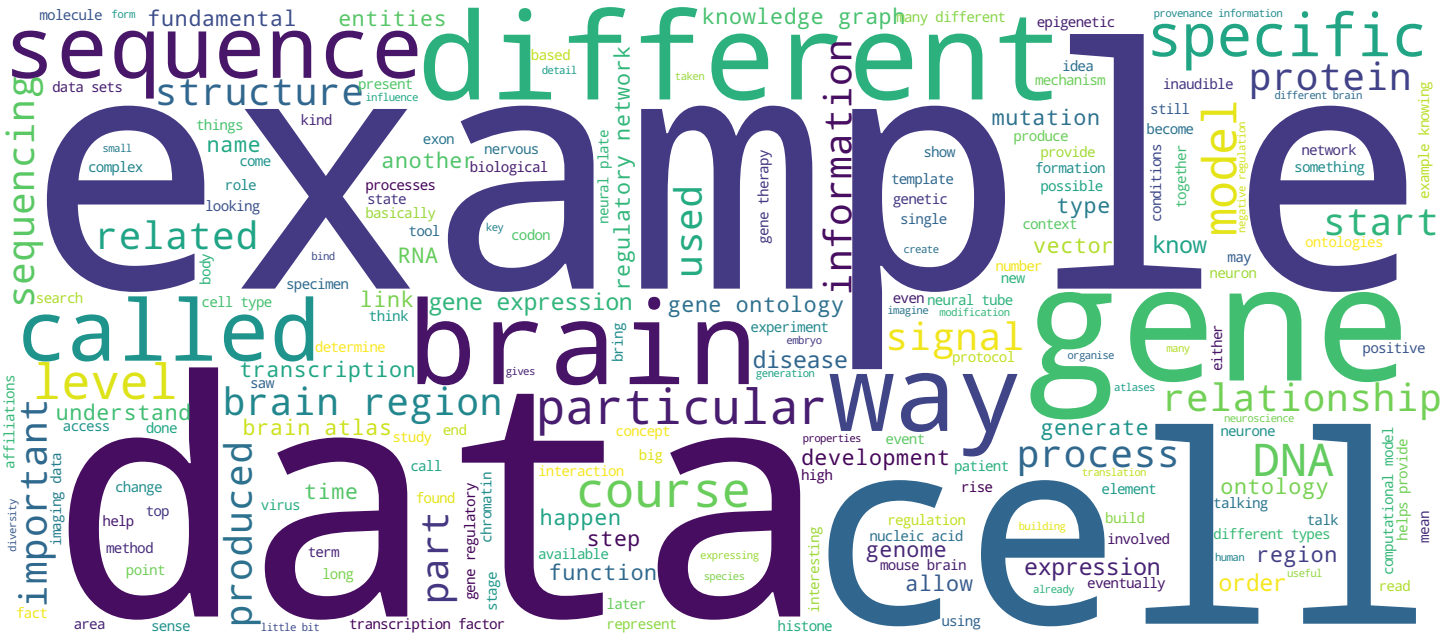




Search MOOC

Video

EPFL

Metadata

Metadata is **data about data**. In order to understand, interpret and use experimental data, one must know:

- Which organism/cell line were used;
- Which conditions and which material were used;
- How the data was normalized and transformed;
- …

Metadata can be descriptive, structural, administrative or statistical. It allows reproducibility, transparency, trust, discovery and attribution of methods, software and data.

Metadata, as I mentioned before, is really data about data. In order to understand, interpret, and use experimental data, we need to know a number of core pieces of information. For example, which organism or cell line was used? What were the conditions under which the experiment performed, and which material or reagents were used? How was the data then normalised or transformed before it was reported? Metadata can be descriptive, it can be structural, it can be administrative in terms of who has access to which data, or it can be statistical. That helps provide enough context to allow for reproducibility, transparency, trust, discovery, attribution of methods, software, and data.

Notes

Summary

0m 05s

Metadata standards

Minimum Information for Neuroscience Datasets (MINDS):

- **Subject**: age, sex, species, strain (when applicable),
- **Methods**: protocols, equipment and parameters used in experiments,
- **Classification**: data category, data format, cell type,
- **Brain location**: ontological term or spatial coordinates/transform,
- **Contributors** and their affiliations,
- **Access** to the data: persistent identifier URL (e.g. DOI)
- **Licence** under which the data is published

One of the things that, if we want to think about searching for data, one of the concepts that our team introduced was the idea of minimum information for neuroscience data sets. MINDS is what we call them. This is really to say, if you want to search for useful data, what are the minimum things you need to know about that data in order to usefully know if you can use that data? For example, knowing properties about the subject, whether it's the age, the sex, the species, or the strain when applicable. What were the methods involved in generating that data, the protocols, the equipment, or the parameters used in experiments? How would you classify that data type? Is it electrophysiology data? Is it brain imaging data? Is it transcriptomic data? What's the binary format, so classification of the type. Where did it come from in the brain? Which brain region? That can be either a semantic name, so a name of a brain region, or it could be actually coordinates in a brain atlas. The contributors and their affiliations. For example, knowing which lab this came from can help you very much understand the type of data it is, the quality, and the affiliations also just to give credit of who actually produced this data.

Notes

Summary

1m 01s

Metadata standards

Minimum Information for Neuroscience Datasets (MINDS):

- **Subject**: age, sex, species, strain (when applicable),
- **Methods**: protocols, equipment and parameters used in experiments,
- **Classification**: data category, data format, cell type,
- **Brain location**: ontological term or spatial coordinates/transform,
- **Contributors** and their affiliations,
- **Access** to the data: persistent identifier URL (e.g. DOI)
- **Licence** under which the data is published

Then is this data actually accessible? Do you have a link? In this case, a persistent identifier, URL, digital object identifiers are ways of getting direct access to the data, and the license under which the data is published. Very often, data may be increasingly open, hopefully, or it may be available in the context of a collaboration. But knowing that upfront helps you understand how much effort is it going to be to actually get access to that data.

Notes

Summary

2m 32s

Metadata organization

Metadata should be **organized** in order to be:

- Machine-accessible
- Searchable
- Linked and integrated

Knowledge graphs connect datasets via their common properties.

Metadata fundamentally should be organised in order to be machine-accessible, should be searchable, should be linked and integrated. Knowledge graphs are a way of organising that to connect data sets via their common properties.

Notes

Summary

3m 07s

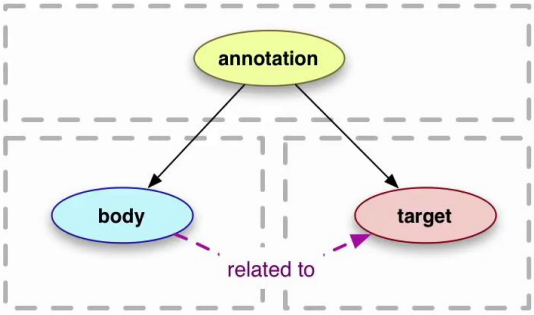
Metadata organization

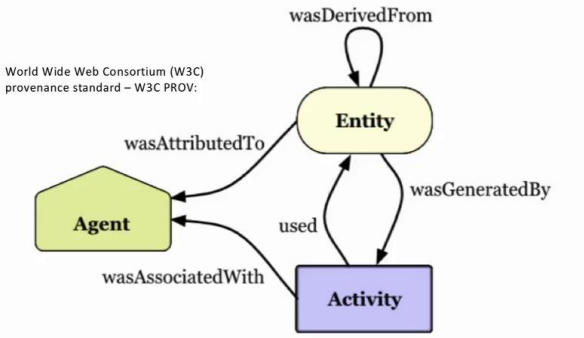
Metadata should be **organized** in order to be:

- Machine-accessible
- Searchable
- Linked and integrated

Knowledge graphs connect datasets via their common properties.

Provenance information provides context and is essential to data reuse





https://www.w3.org/TR/prov-primer/

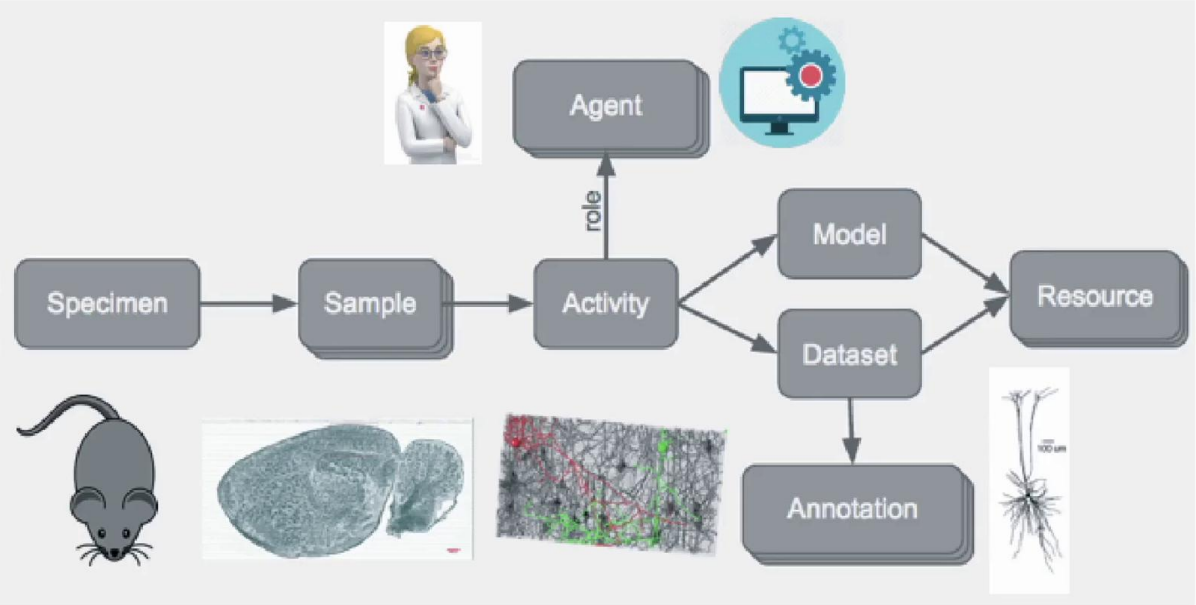
In this case, when we're talking about annotations, as we were earlier, you would want to know, for example, how is the annotation related to the body of the text and related to, for example, the brain region of interest? Knowledge graphs give you a way of expressing those different entities and their relationships in a useful and flexible way. One of the additional things that knowledge graphs allow you to do is include provenance information. For example, knowing how was that data produced? For example, who produced it? How was it produced? Under which conditions, which protocols? Was the data analysed using a particular software? Who were the people involved? All of that together really helps provide a rich description of the data and the context within it which it was produced.

Notes

Summary

3m 26s

Provenance graph example



This is an example to show how you could represent some provenance information in a knowledge graph with describing the specimen, how the sample was taken from the specimen, what was the activity used to produce that sample? By whom? The agent? Was the individual, person, that a model was produced from a given data set? All of that ends up in, for example, a model neuron through this process that gives a richer description of the process and the steps and who was involved in producing the data.

Notes

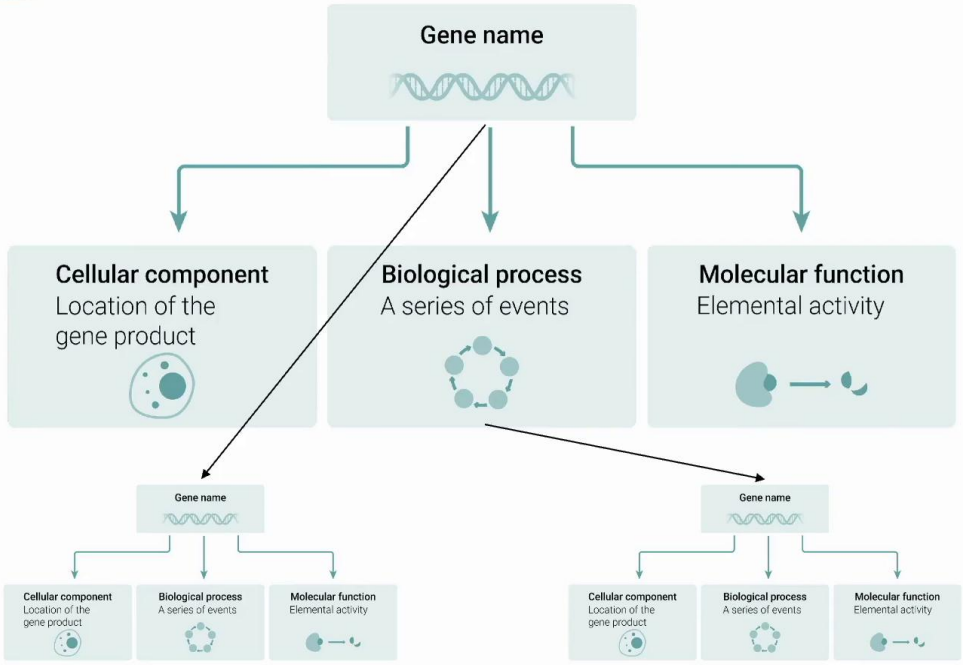
Summary

4m 25s

Example: GO

Components:

- Gene name, product name and GO ID
- Cellular location of product and GO ID
- Molecular function(s) of product and GO ID(s)
- Biological process(es) product is involved in and GO ID(s)
- Relationship to other terms



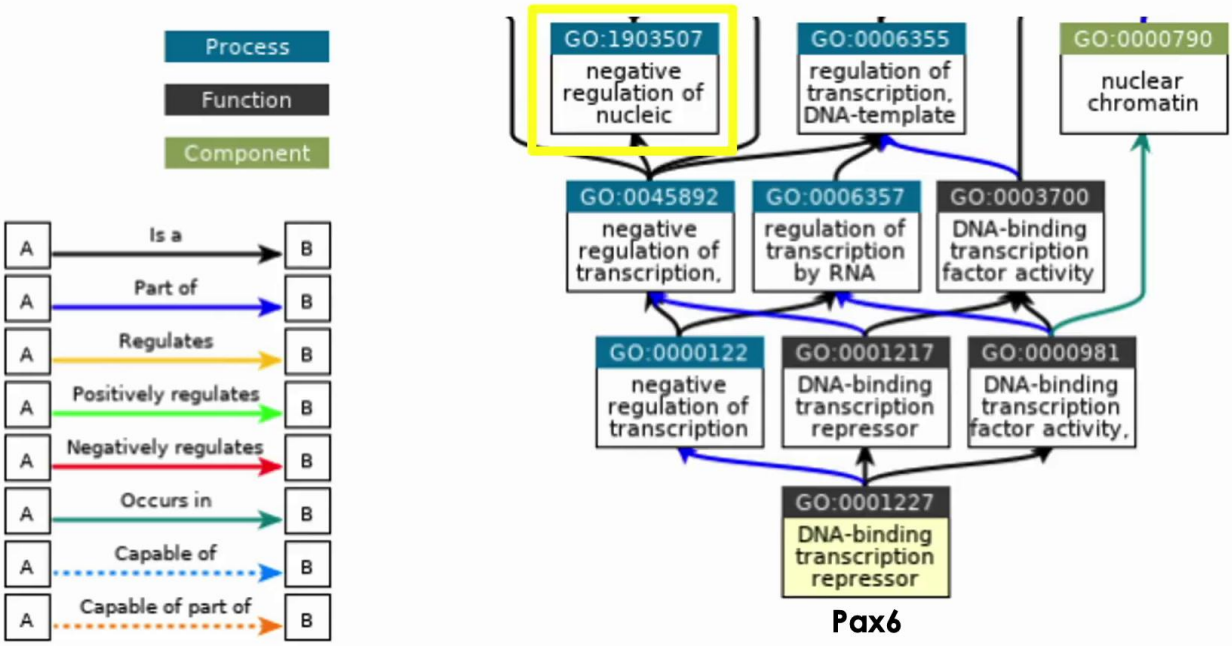
An ontology is a more formal representation of knowledge about a specific set of entities. It can provide a controlled vocabulary for talking about a specific set of entities or concepts, and the properties of these entities, as well as the relationship between these entities. An example is the GO ontology, the gene ontology. That includes things like the gene name, the GO ID, the gene ontology identifier, information such as the relationship to a cellular component, which is the location of where that gene product was actually found, the molecular function, so it describes different categories of elemental activities in the physiological network. Those processes, the series of events that that element, that entity is involved in. Of course, the relationship to the rest of the genes, and processes, and functions, and cellular locations throughout this complex biophysical system.

Notes

Summary

5m 07s

GO – Pax6

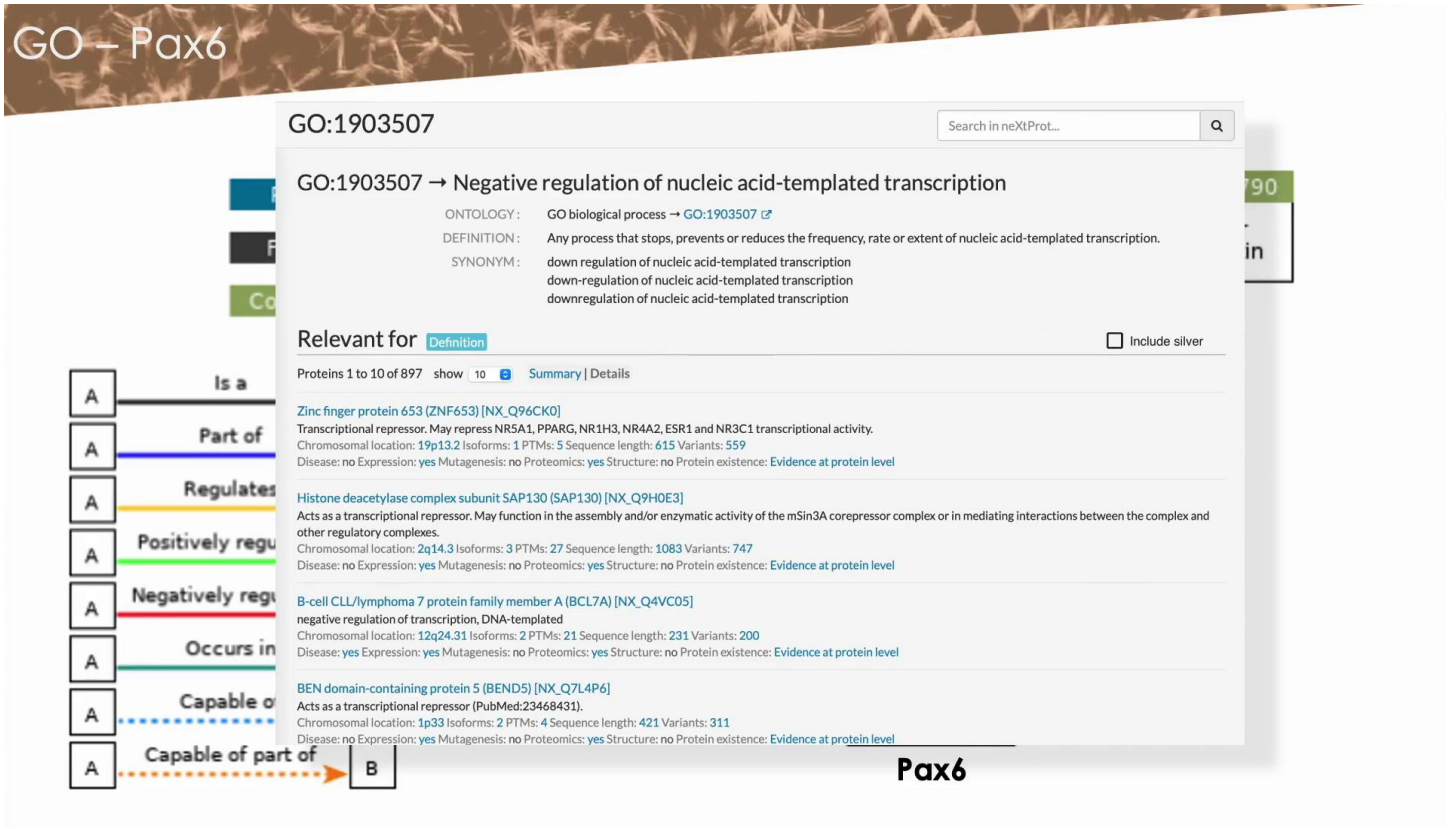


Here's an example where we're looking at PAK6, which is actually a gene related to the production of PAK6 proteins that is related to organ development, for example, particularly in the olfactory bulb and in the eye. Its expression is actually controlled by a complex regulatory network. There's many different types of relationships. For example, it can be part of a transcription process. There can be positive and negative regulation. There can be different locations where those elements occur. If we take one of these elements of this network, for example, this GO:1903507, we can actually take that identifier and look it up in the knowledge graph of the gene ontology and find out that it's actually related to the negative regulation of nucleic acid templated transcription.

Notes

Summary

6m 22s

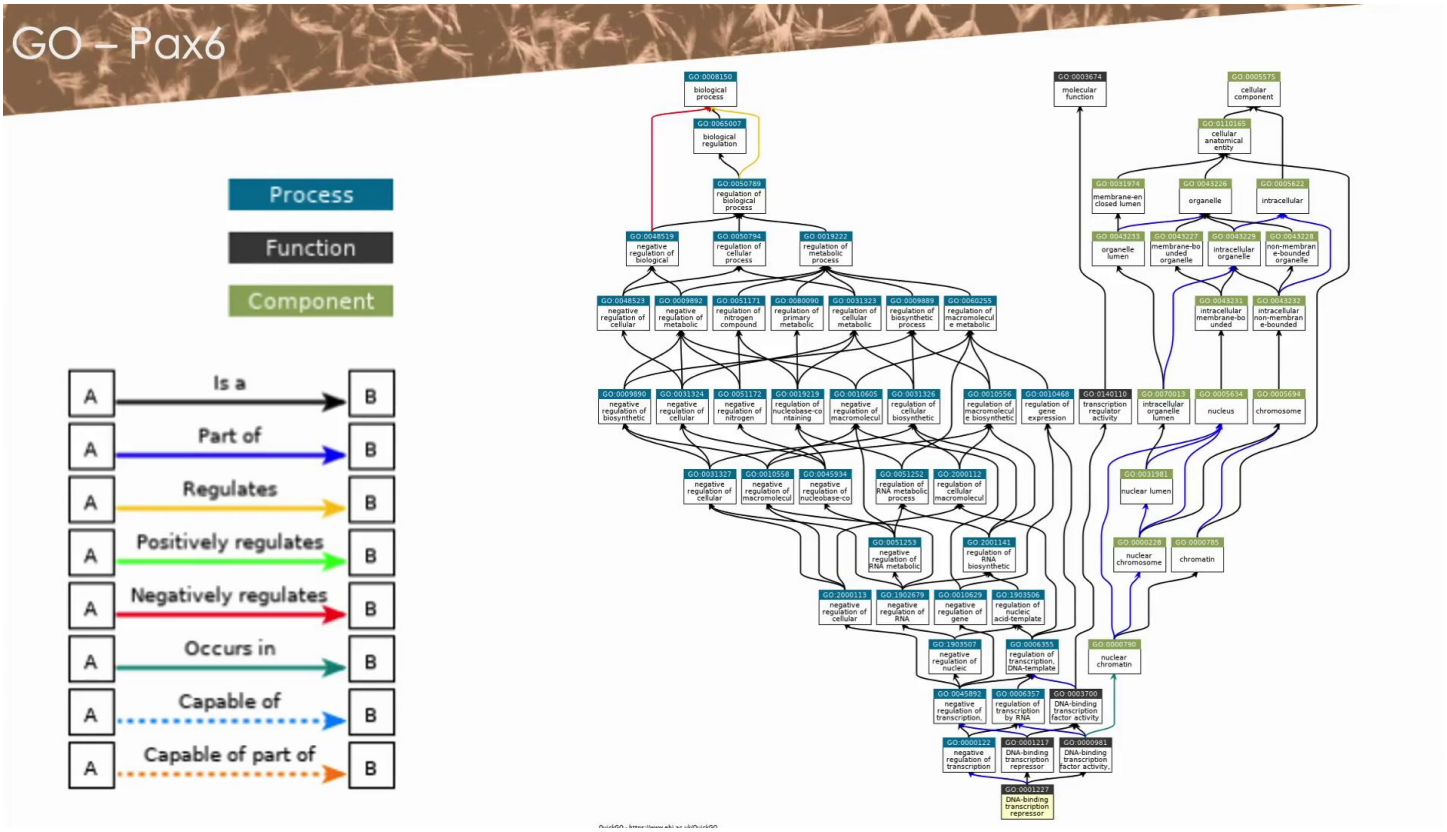


There's specific information, including the definition, these other known synonyms, and then the fact that it's related to 897 other entities within the gene ontology.

Notes

Summary

7m 26s



As you can see, the gene ontology, as it accumulates data from many, many studies, it actually builds up and organises quite a large amount of information about the very complex biochemical system that underlies our cellular processes.

Notes

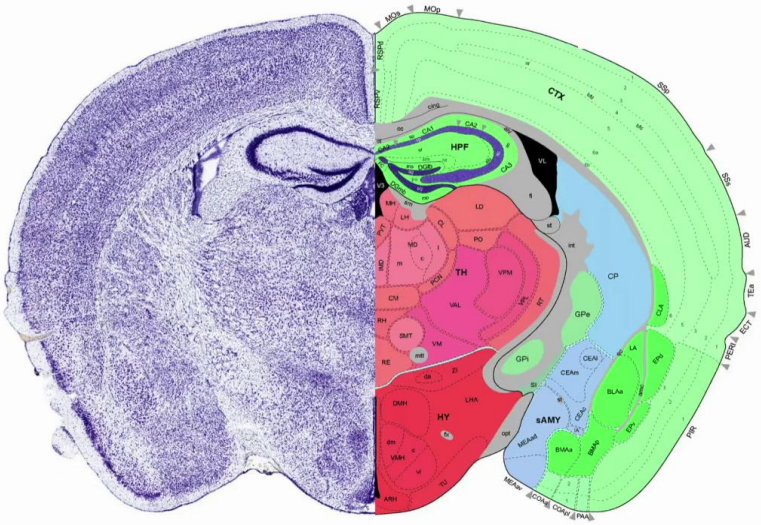
Summary

7m 42s

Example: brain region ontology

Components:

- Anatomical boundaries
- Landmarks
- Standard name and abbreviation
- Relationship to parent and substructures



Mouse atlas data courtesy of the Allen Institute for Brain Science

Now, a brain ontology is another type of an ontology. This is a way of providing standardised naming, for example, for anatomical boundaries, for landmarks, giving them a standard name and abbreviation, and of course, the relationship between these different brain regions.

Notes

Summary

8m 03s

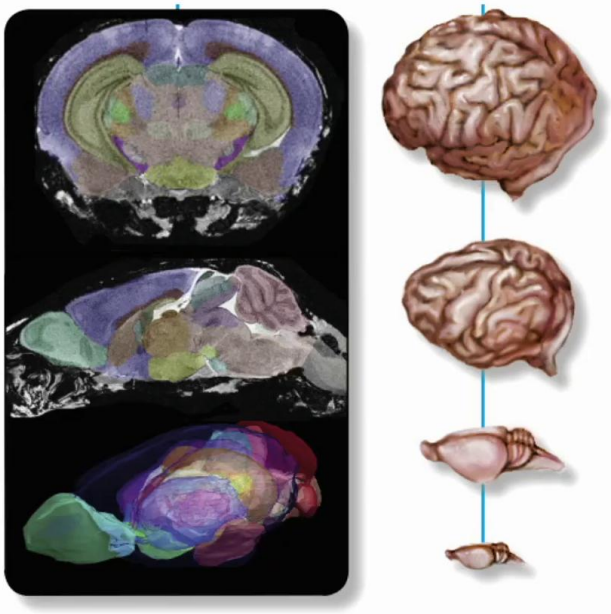
Brain atlases

Brain atlases establish a **link between semantics and space**. The three main constituents of a brain atlas are:

- A coordinate space
- A parcellation showing the 3D boundaries of parcels
- Ontology of brain regions

They can be used to organize data and link it to:

- Gene expression data
- Neuronal types
- Imaging data
- Physiology
- Models



Mouse atlas data courtesy of the Allen Institute for Brain Science

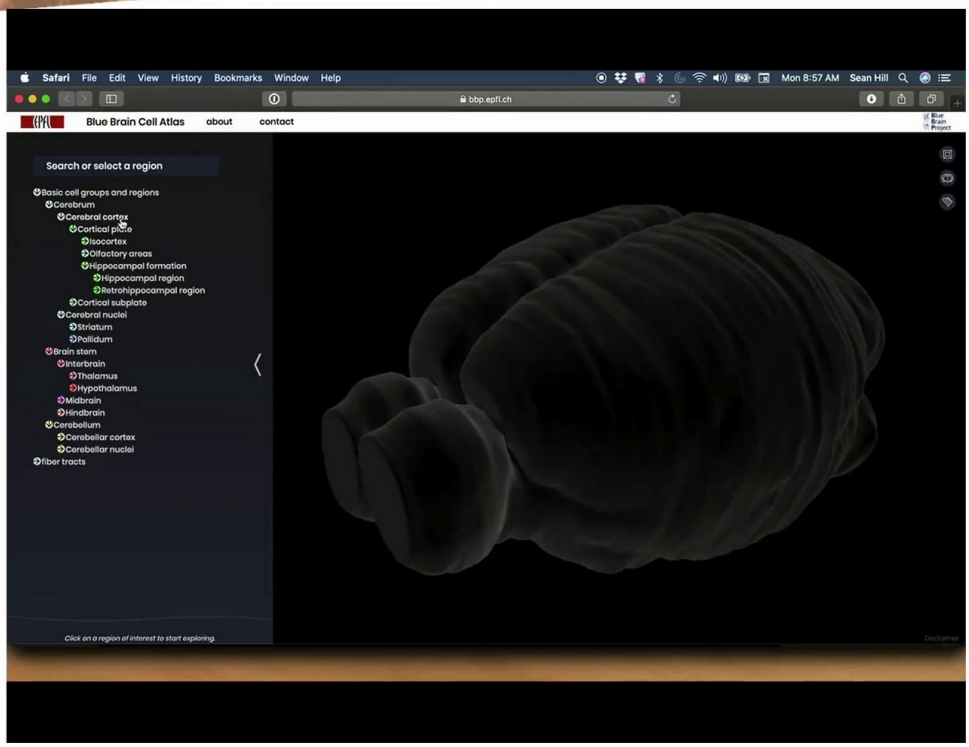
Brain atlas actually establish a link between the names of the regions and their actual location. The three main constituents of a brain atlas are a coordinate space, a parcelation showing the 2D or 3D boundaries of these parcelations in precise coordinates, and then an ontology of brain regions, which is their names and their relationships between the different brain regions. Those can be used to organise data and to link to many different types. For example, gene expression data and neuronal types, imaging data, physiological data, and computational models. It provides a very useful framework for saying where in the brain are you talking about? Where in the brain did this data come from? Where in the brain are you actually building a model of? How we can bring all of these elements together, the ontologies, the atlases, the data, and the computational modeling?

Notes

Summary

8m 23s

Atlases – multiscale & multimodal



https://bbp.epfl.ch/nexus/cell-atlas/

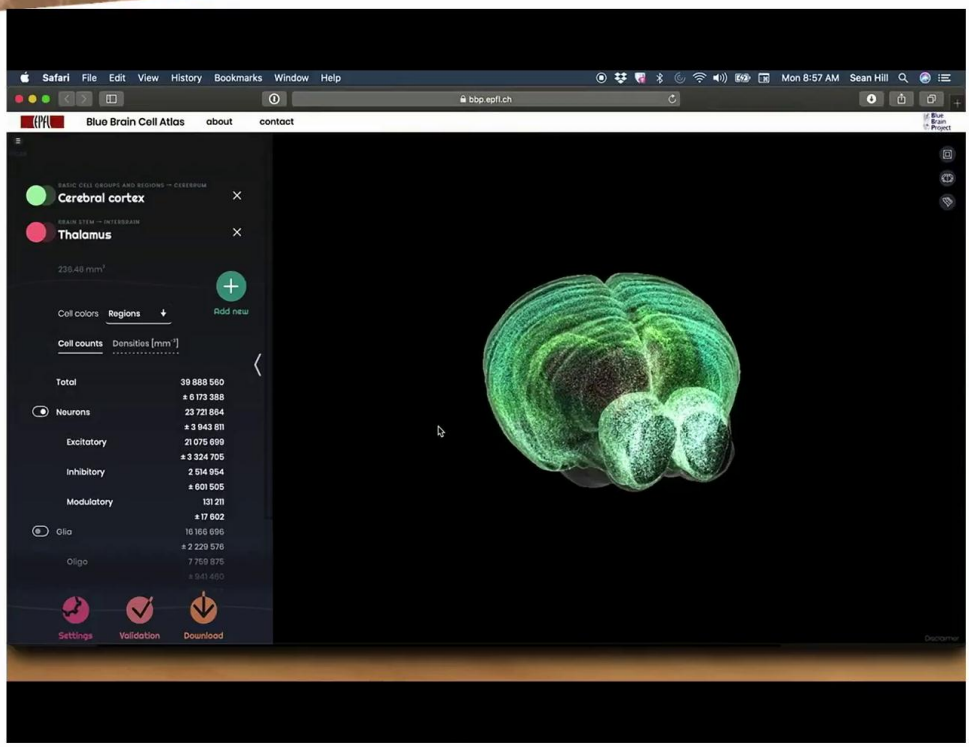
It's actually exemplified through the blue brain cell atlas.

Notes

Summary

9m 28s

Atlases – multiscale & multimodal



https://bbp.epfl.ch/nexus/cell-atlas/

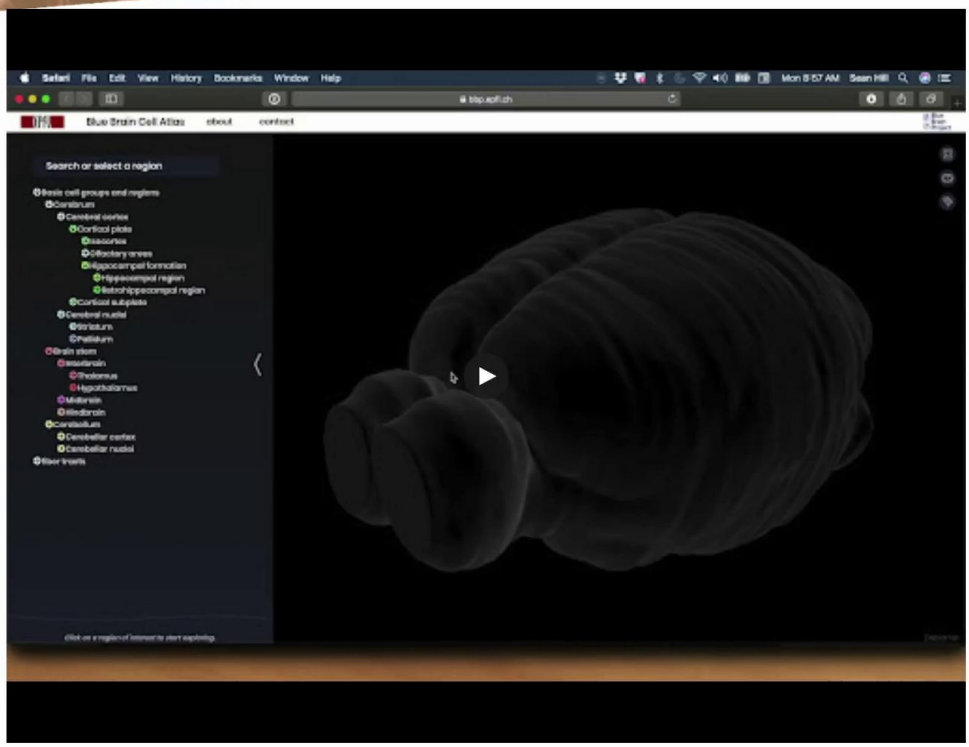
Here, we use the ontologies, the brain regions, the brain atlas, to organise the outputs of a computational model that predicts cell positions for all the cells in the mouse brain.

Notes

Summary

9m 33s

Atlases – multiscale & multimodal



https://bbp.epfl.ch/nexus/cell-atlas/

This is how we can start to organise additional layers of data in order to really facilitate larger and more sophisticated computational models.

Notes

Summary

9m 45s