Source Data – Making Data Discoverable

Sara El-Gebali [1], Nancy George [1], Robin Liechti [2], Lou Götz [2], Isaac Crespo [2], Ioannis Xenarios [2], Thomas Lemberger [1]

1-EMBO, Heidelberg, Germany, sourcedata@embo.org
2-Vital-IT, SIB Swiss Institute of Bioinformatics, Switzerland

Most research journals publish data as figures which are not machine-readable. The SourceData project aims at providing new ways for scientists to efficiently find research data in the published literature. We develop tools to make the data underlying published papers easier to find, to promote their reuse and to connect data across the scientific literature.

Descriptive metadata

Published experiments are described using recognized standardized terminologies to reduce ambiguities. In a second step, entities are assigned to experimental "roles" to capture the essence of the experimental design. SourceData is developing a user-friendly web-based curation platform that helps data editors and biocurators to perform these key tasks.

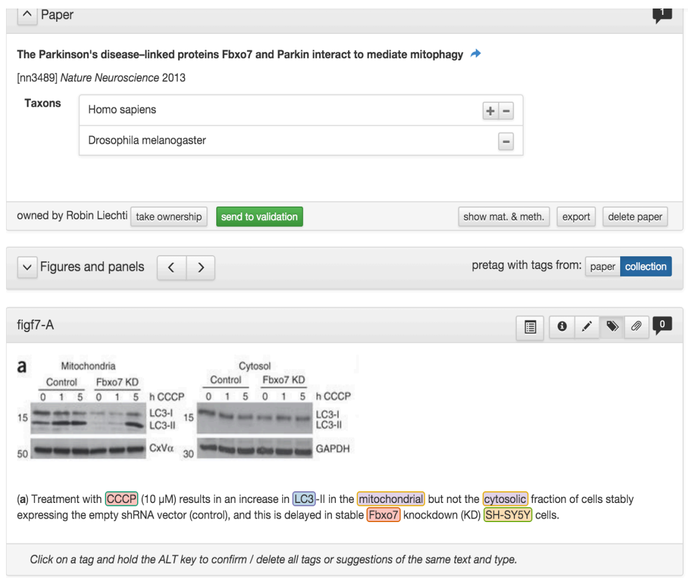
Entities and resources

Public Resources	Entity Type
ChEBI/PubChem	Small Molecules
NCBI Gene/Rfam	Genes
Uniprot	Proteins
Gene Ontology/ Corum	Subcellular location
Cello-saurus	Cell type
Uberon	Tissue
NCBI Taxonomy	Organism
No external resource	Unidentified

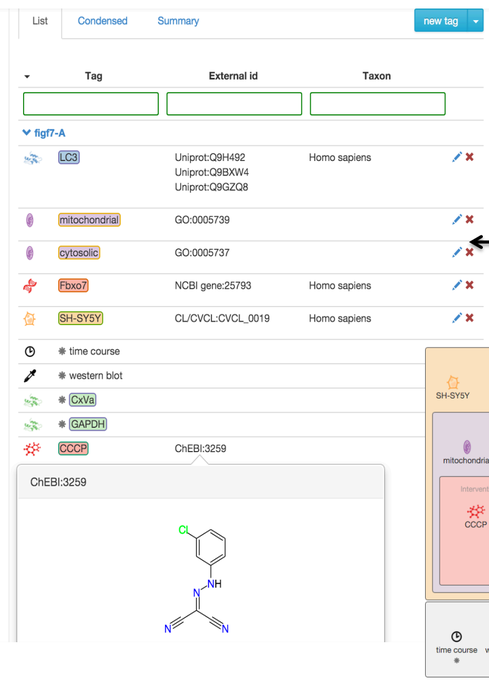
Entity Role

- what is being actively changed to test a **causal hypothesis**?
- what is being **measured**?
- what is used as a **baseline** measurement?
- what is used as a **proxy m**easurem-ent?
- what distinguishes **experimental groups**?
- any other entities

Tagging of figures and their legends



Integrated curation interface



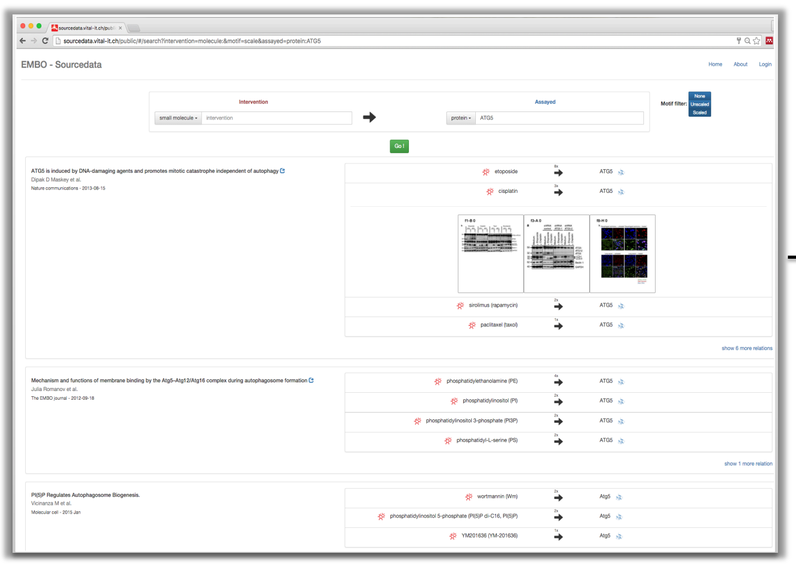
List of identifiers

Visual summary

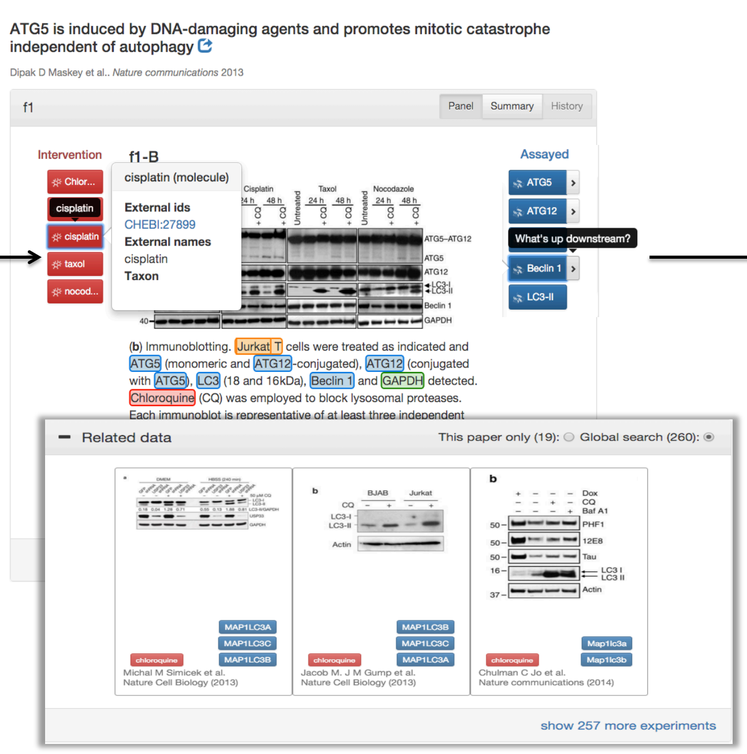
Data-oriented search

Searching and navigating the 'knowledge graph' built from connected data will open new possibilities for scientists to browse, search and mine the scientific literature in a more efficient and intuitive manner. Combining human-readable figures and machine-readable descriptive metadata into 'smart figures' overcomes the limitations and complements the current keyword search methods.

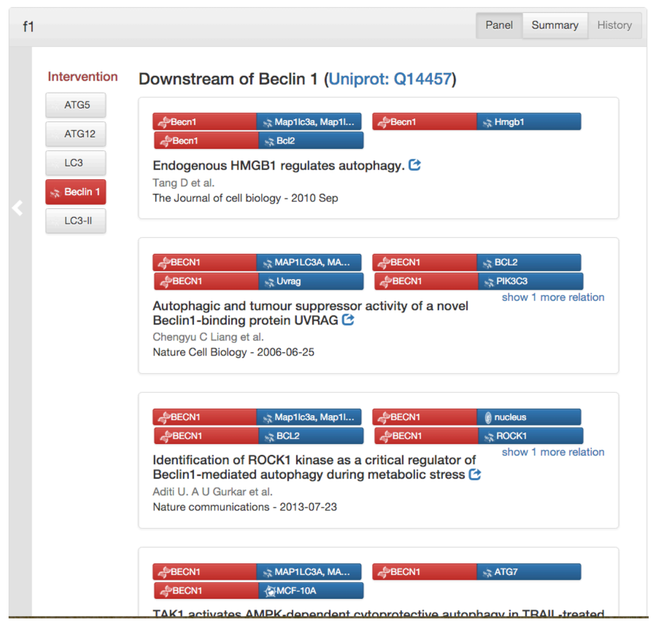
Search interface



SmartFigure to visualize data in context

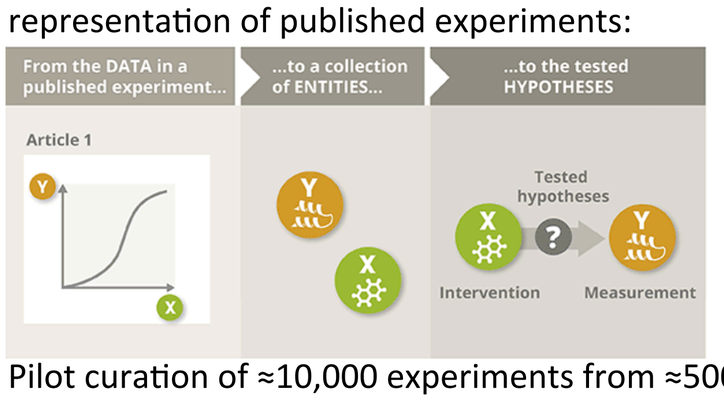


Browsing by traversing the knowledge graph



Summary

The SourceData curation tool uses a two step process to extract a machine-readable representation of published experiments:



Pilot curation of ≈10,000 experiments from ≈500 papers related to the field autophagy shows that our model can be used in 80% of the cases.

The resulting knowledge graph comprises ≈30,000 hypotheses directly linked to the respective experimental evidence.

Future Direction

Integrating SourceData in the EMBOpress publishing workflow:
The curation of live manuscripts is currently being tested at EMBOpress journals:

Pre-acceptance → Curation by data editor → Authors' approval → Publish

Improve automation:
Text-mining and machine learning will be progressively integrated to automate repetitive tasks and speed up curation.

Figures as stuctured data + metadata packages
Figures associated with tabular source data will be combined with machine-readable metadata so that they can be exchanged as integrated machine- and human-readable packages.

Sign up as beta-tester at:

http://sourcedata.embo.org

Our Partners

Robert Bosch Stiftung

EMBOpress HighWire

WILEY

SIB Swiss Institute of Bioinformatics