SourceData: Making Data Discoverable

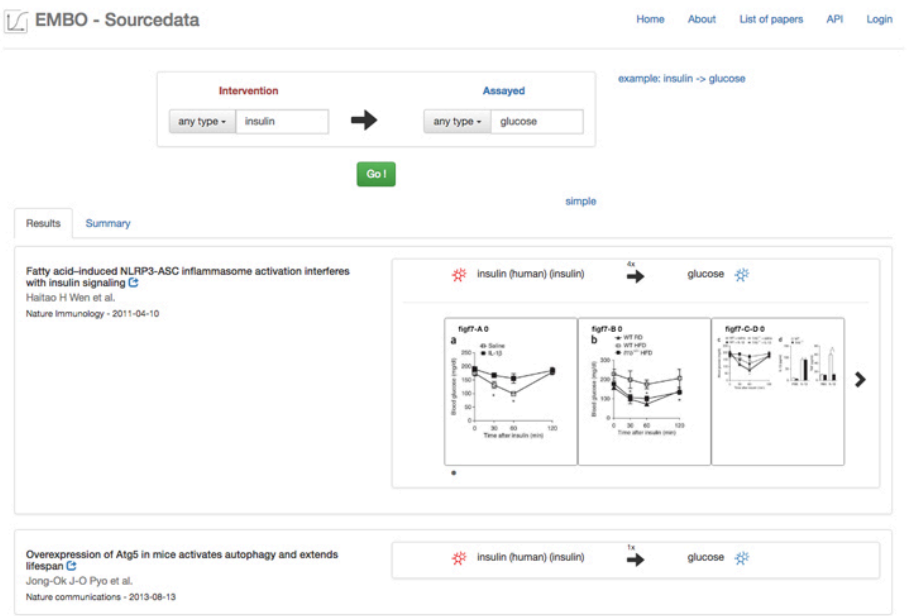
Nancy George[2,*], Robin Liechti[1,*], Sara El-Gebali[2,*], Lou Götz[1,*], Isaac Crespo[1], Ioannis Xenarios[1], Thomas Lemberger[1,3]

*contributed equally,
[1] Vital-IT, Swiss Institute of Bioninformatics, Lausanne Switzerland
[2] SourceData, EMBO, Heidelberg, Germany
[3] Correspondence to: thomas.lemberger@embo.org)

In molecular and cell biology, most of the data presented in published papers are not available in formats that allow for direct analysis and systematic mining. The goal of the SourceData project (http://sourcedata.embo.org) is to make published data easier to find, to connect papers containing related information and to promote the reuse and novel analysis of published data. The main concept underlying the project is that the structure of a dataset provides information about the design of the study in question and can be exploited in powerful data-oriented search strategies. SourceData has therefore developed tools to generate machine-readable descriptive metadata from figures in published manuscripts. Experimentally tested hypotheses are represented as directed relationships between standardized biological entities. Once processed, a comprehensive 'scientific knowledge graph' can be generated from this data (see demo video1 at https://vimeo.com/sourcedata/kg), making the body of data efficiently searchable. Importantly, this graph is objectively grounded in published data and not on the potentially subjective interpretation of the results.

SourceData has developed algorithms to efficiently search the data-oriented knowledge graph and an interface, shown in Figure 1, that enables users to find paper based on their data content:

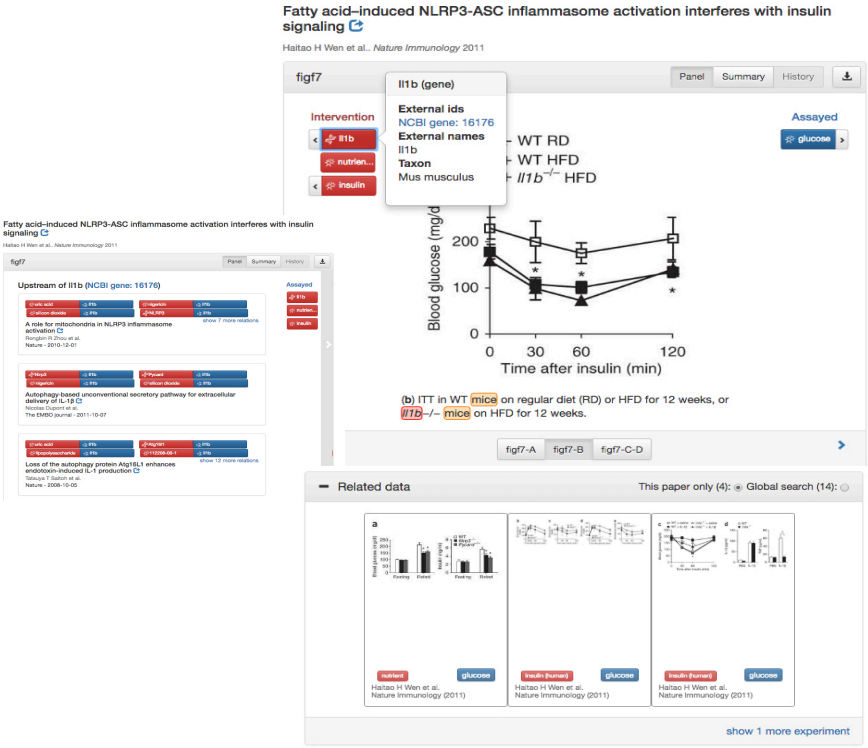
Figure 1. SourceData Search interface



The search capabilities have also been incorporated into the SmartFigure viewer. This application can be embedded directly into online publications and allows the

visualisation of figures in the context of related data published elsewhere. Readers can then navigate from one figure to the next by following linked entities (see Figure 2 and see demo video2 at http://vimeo.com/sourcedata/search).

Figure 2. 'SmartFigures" viewer



Access to the SourceData database by computer programs is provided through a public Application Program Interface (http://sourcedata.vital-it.ch/public/#/api), giving developers the chance to produce their own software solutions or machine-driven analyses based on the SourceData data format.

Future perspectives for the project include the integration of a structured representation of time and incorporating descriptions of experimental procedures and reagents. Furthermore, SourceData will develop portable 'figure/data packages' that combine and cross-link the human-interpretable figure to the underlying machine-readable metadata and data files. This means linking the original experimental dataset with the representative experimental figure to allow ease of re-use and transparency of data. Finally, we plan to adapt the SourceData model to integrate existing approaches for the representation of large-scale biological data.

With SourceData, we are developing a platform that simultaneously improves the discoverability and utility of research data and of the scientific articles where these data are reported. It will therefore provide the basis for a reward system that will incentivize authors to share their data openly, thus driving a broader adoption of open data and open science by the community.