The ever-increasing availability of data and processing resources in the area of text mining has made possible the construction of large knowledge bases from the scientific literature. In many practical applications, however, this process cannot be fully automated due to the complexity of the data and the performance of available tools, so that expert user intervention is still required for validating the extracted data.

The BioHub Information and Knowledge Management System (IKMS) seeks to gather knowledge about biorenewable feedstocks as sources for chemical ingredients for the manufacture of surfactants. It uses a semi-automated process of ontology construction by utilising a range of semantic technologies including text mining, OWL, SPARQL and JSON. These tools are integrated within a bottom-up, data-driven curation architecture organised in three separate layers.

The bottom layer is a text mining layer that extracts information about sustainable feedstocks and their component chemical substances from scientific documents and stores it in an OWL ontology.

In an intermediate layer, candidate assertions of the relationships between entities of interest are selected for curation via SPARQL queries to the text mining output. An assertion generated at this stage is converted from OWL to a JSON object for input to a curation layer. Assertions may include, for instance, extracted facts about chemicals and associated feedstocks, transformations reported to extract a particular chemical from a feedstock, statements about a chemical and its surfactant properties, and so on.

A final curation layer is used to let curators browse, edit and validate the candidate assertions. Curation is implemented via a Web-based interface where assertions are presented in tabular format linked to the parts of documents in which they are mentioned (sentences or paragraphs). The results of the curation phase are validated OWL classes and axioms about feedstocks and chemicals which are stored in the BioHub knowledge repository and can be passed to subsequent processes for knowledge discovery and innovations in the IKMS.