Text-mining needs of the food microbiology research community

Estelle Chaix, Sophie Aubin,
Louise Deléger and Claire Nédellec
Speaker: Robert Bossy

IN-OVIVE 2017 at EFITA WCCA congress - July 2nd - 6th - Montpellier

Context

Microorganisms : very abundant and can live in extreme conditions

Microbial diversity research:

- Microbiomes
- Microbial interactions and ecosystems
- Phylogeny

For :

- Human, plant or animal health
- Plant growth, bioremediation or food processing

Use molecular technologies (DNA-sequencing and metagenomics)

openMINTED

Context

Data and knowledge sources on microbial biodiversity:

- Experimental data
- Curated databases
- Textual documents : scientific publications, reports, patents or medical records

From Microbial biodiversity community needs
To Ontology-based Text-Mining applications

Context

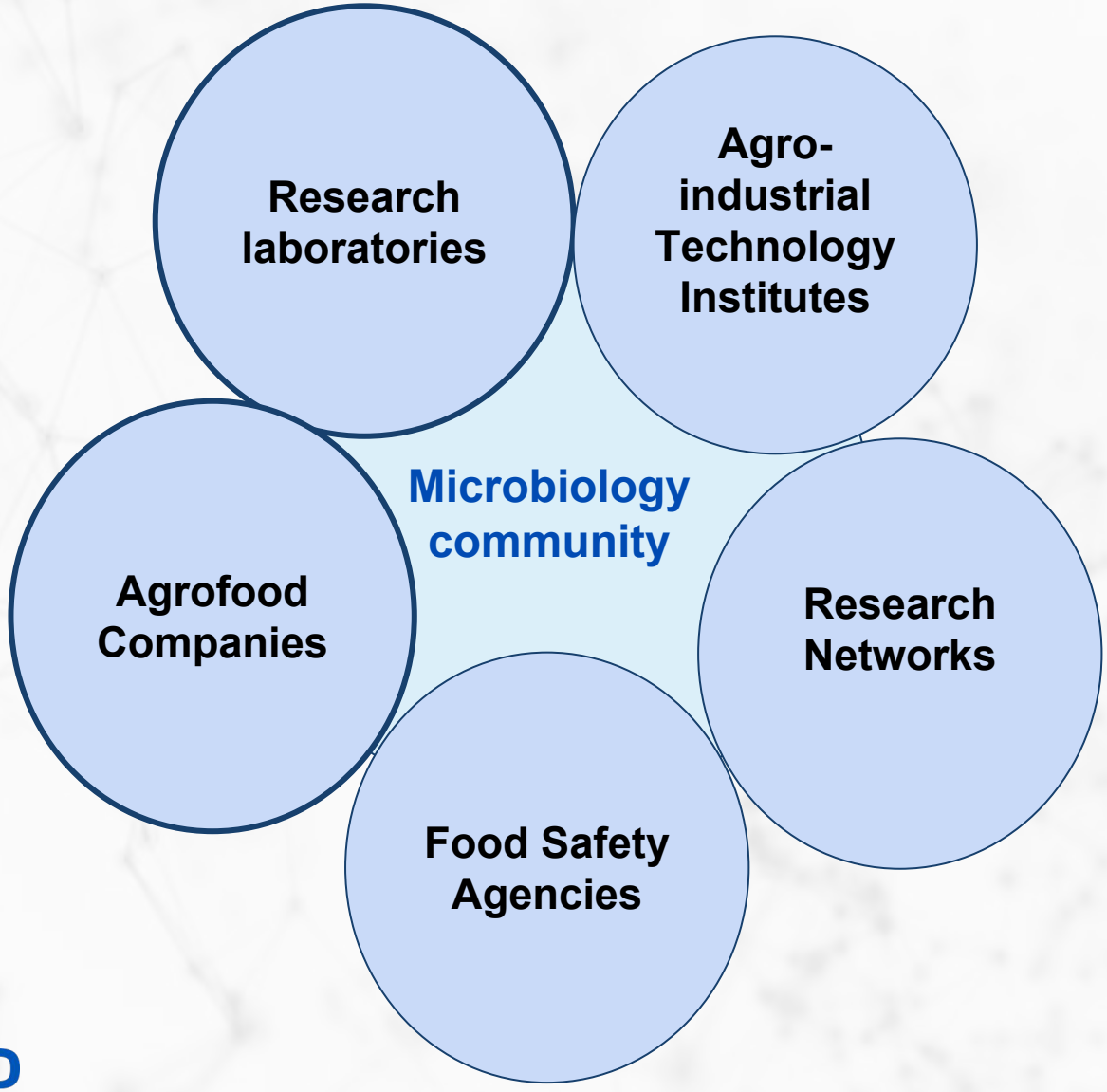
This work is part of the development of the European infrastructure for text-mining OpenMinTeD (http://openminted.eu/)

This project targets 4 different research communities through different use-case examples. This work focuses on the Agriculture and Biodiversity field.

Methodology of user needs analysis

1. Identification of end-users and preliminary needs → scope of the use-case
2. Identification of stakeholders, their needs and theirs interactions
3. Validation of the requirements provided by the need analysis
4. Design of the TDM solution

1. End-user identification



2. End-user needs

Target: sub-community of researchers working on positive microbial flora

Method : User-centered method (persona)

1. Who is the persona?
2. What information are they interested in?
3. How do they need to access the information?
4. What solutions would satisfy their needs?

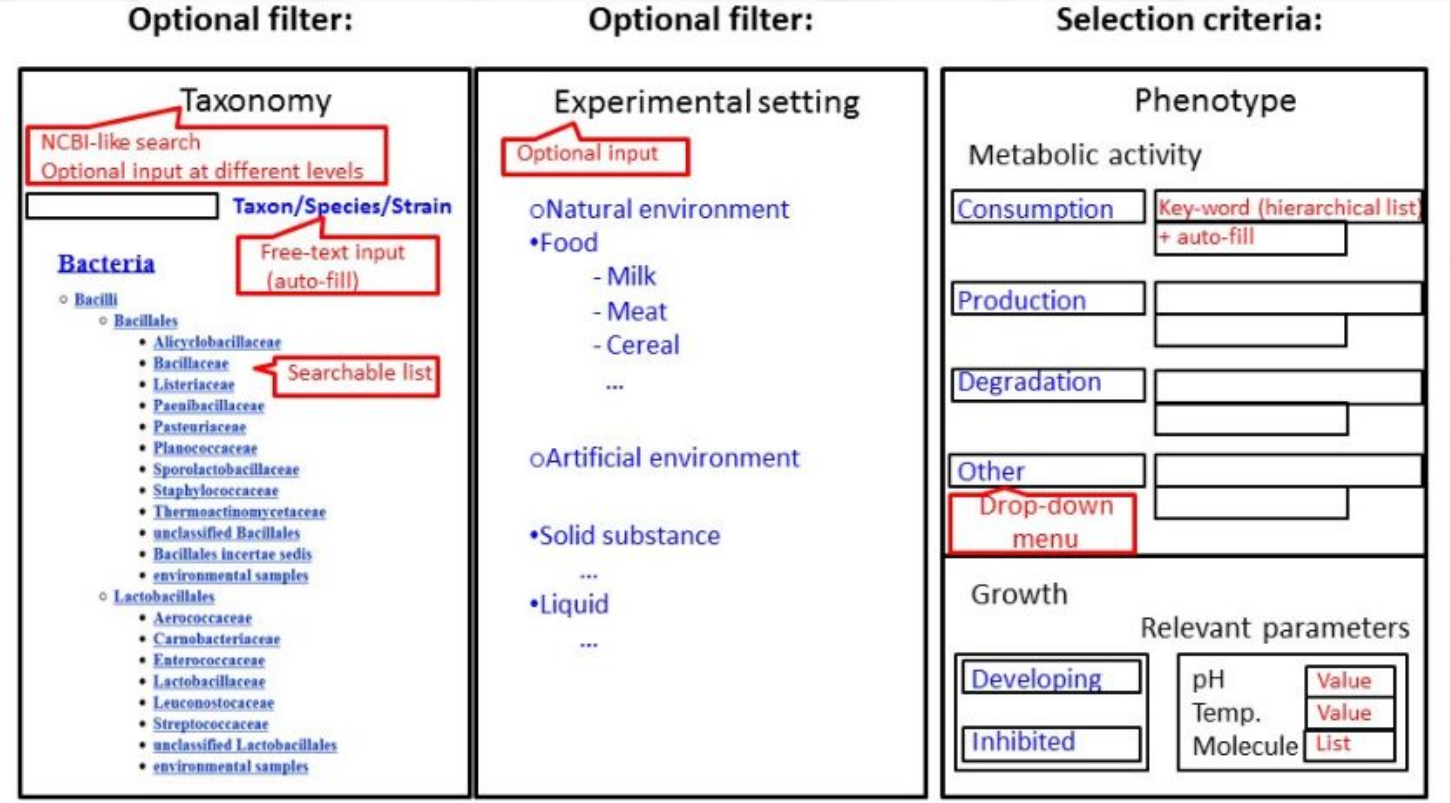
2. Persona example

Lily: 38 years old, former baker and expert in bread leavens

- She wants to compare commercial and artisanal leavens.
- She needs to know the habitats and phenotypes of bread starter cultures.
- She reads scientific papers, browse microbial collections, draws expertise from experience and baker networks.
- Requires: data and knowledge about microbial phenotypes and taxonomic diversity of leaven.

2. End-user interface

One main interface for all personas needs



3. Validation of the end-user needs

- Microbial biodiversity: qualify microorganism provenance (prevent contamination).
- Microbial phenotype: degradation and production of molecules.
- Industrial use: impact of microbial phenotype in the context of food processing (flavor, taste, biopreservation).

A common vocabulary of microbe habitats and phenotypes.

4. TDM solutions

Extraction from scientific publications

- relevant entities (microorganisms, habitats, molecules, phenotypes and applications) and their normalization with ontologies.
- Relations between entities.

State-of-the-art resources

- Bacteria Biotope task of the BioNLP-Shared Task for microorganisms and habitats.
- OntoBiotope ontology for habitats and phenotypes.
- BioCreative IV CHEMDNER Task for molecules.
- Methods for relation extraction: TEES or AlvisRE.

Conclusion

- Converging needs identified from diverse fictional users.
- TDM solutions have to be linked to traditional data sources.
- Ontologies as a means to aggregate different sources.