

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)



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

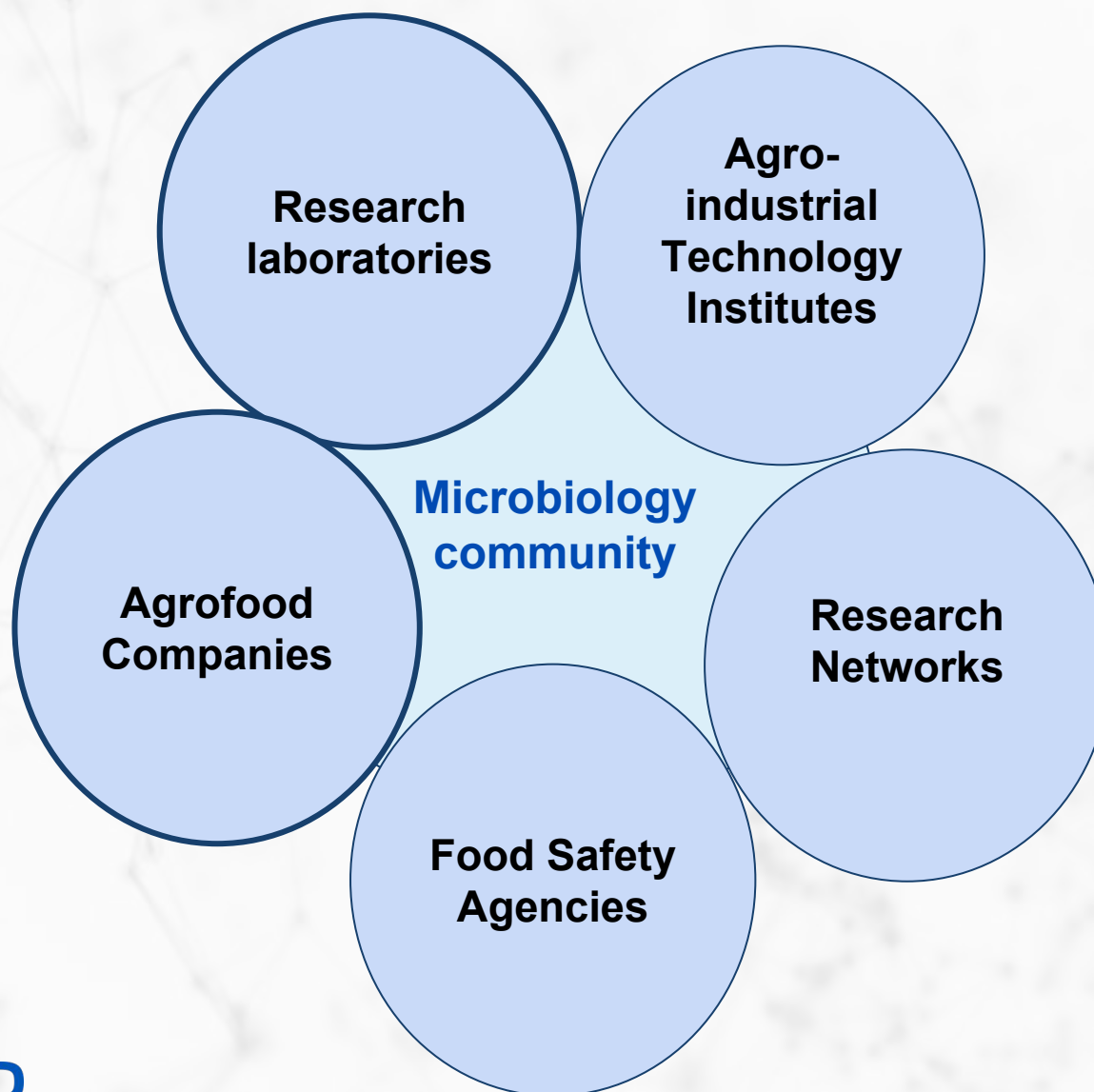
This work is part of the development of the European infrastructure for text-mining OpenMinTeD (<http://openminted.eu/>)

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

Optional filter:	Optional filter:	Selection criteria:																
<p data-bbox="392 579 546 612">Taxonomy</p> <p data-bbox="247 625 581 676">NCBI-like search Optional input at different levels</p> <p data-bbox="446 689 668 711">Taxon/Species/Strain</p> <p data-bbox="266 743 369 765">Bacteria</p> <ul data-bbox="272 782 542 1265" style="list-style-type: none">o Bacilli<ul data-bbox="311 803 542 1075" style="list-style-type: none">o Bacillales<ul data-bbox="349 825 542 1075" style="list-style-type: none">• Alicyclobacillaceae• Bacillaceae• Listeriaceae• Paenibacillaceae• Pasteuriaceae• Planococcaceae• Sporolactobacillaceae• Staphylococcaceae• Thermoactinomyetaceae• unclassified Bacillales• Bacillales incertae sedis• environmental sampleso Lactobacillales<ul data-bbox="349 1096 542 1265" style="list-style-type: none">• Aerococcaceae• Carnobacteriaceae• Enterococcaceae• Lactobacillaceae• Leuconostocaceae• Streptococcaceae• unclassified Lactobacillales• environmental samples <p data-bbox="469 732 624 783">Free-text input (auto-fill)</p> <p data-bbox="498 853 653 875">Searchable list</p>	<p data-bbox="768 579 1078 612">Experimental setting</p> <p data-bbox="730 632 904 654">Optional input</p> <ul data-bbox="745 689 1006 1143" style="list-style-type: none">o Natural environment<ul data-bbox="745 725 1006 886" style="list-style-type: none">• Food<ul data-bbox="803 753 890 851" style="list-style-type: none">- Milk- Meat- Cereal...o Artificial environment<ul data-bbox="745 1015 938 1143" style="list-style-type: none">• Solid substance...• Liquid...	<p data-bbox="1354 579 1518 612">Phenotype</p> <p data-bbox="1184 632 1425 661">Metabolic activity</p> <table border="1" data-bbox="1170 682 1647 1043"><tr><td data-bbox="1170 682 1363 715">Consumption</td><td data-bbox="1373 682 1647 743">Key-word (hierarchical list) + auto-fill</td></tr><tr><td data-bbox="1170 772 1363 805">Production</td><td data-bbox="1373 772 1647 833"></td></tr><tr><td data-bbox="1170 862 1363 895">Degradation</td><td data-bbox="1373 862 1647 923"></td></tr><tr><td data-bbox="1170 952 1363 985">Other</td><td data-bbox="1373 952 1647 1013"></td></tr></table> <p data-bbox="1193 986 1348 1038">Drop-down menu</p> <p data-bbox="1184 1072 1290 1100">Growth</p> <table border="1" data-bbox="1170 1115 1647 1265"><thead><tr><th colspan="2" data-bbox="1354 1108 1634 1136">Relevant parameters</th></tr></thead><tbody><tr><td data-bbox="1170 1150 1363 1183">Developing</td><td data-bbox="1402 1150 1634 1183">pH Value</td></tr><tr><td data-bbox="1170 1212 1363 1245">Inhibited</td><td data-bbox="1402 1212 1634 1245">Temp. Value</td></tr><tr><td></td><td data-bbox="1402 1219 1634 1252">Molecule List</td></tr></tbody></table>	Consumption	Key-word (hierarchical list) + auto-fill	Production		Degradation		Other		Relevant parameters		Developing	pH Value	Inhibited	Temp. Value		Molecule List
Consumption	Key-word (hierarchical list) + auto-fill																	
Production																		
Degradation																		
Other																		
Relevant parameters																		
Developing	pH Value																	
Inhibited	Temp. Value																	
	Molecule List																	

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.