# Text-mining needs of the food microbiology research community

Estelle Chaix, Sophie Aubin, Louise Deleger, Claire Nédellec

# Text-mining needs of the food microbiology research community

Estelle Chaix[1], Sophie Aubin[2], Louise Deléger[1], and Claire Nédellec[1]

[1] MaIAGE, INRA, Université Paris-Saclay, 78350 Jouy-en-Josas, France
[2] DIST, INRA RD 10, Route de Saint-Cyr, Université Paris-Saclay, 78026 Versailles CEDEX, France
firstname.lastname@inra.fr

**Abstract.** To ensure the usefulness of a bioinformatics service, analysis of user needs is an essential step. Furthermore, if the service anticipates the identified needs, acceptance by the user is easier. The aim of this work is to provide an overview of the requirements of a microbial diversity research community for ontology-based text-mining applications. This study is part of the development of the European infrastructure for text-mining, OpenMinTeD, that targets Biodiversity among other research fields. The requirement analysis was completed through targeted online surveys, interviews, focus group meetings and workshops. This work yields to a detailed up-to-date landscape of stakeholders (data provider, producer and consumer), their potential role and their expectations of general interest with respect to text-mining applications. We introduce a user-centered approach to focus on microbiologist end-user functional requirements, including application user interfaces. The resulting description of these needs guides OpenMinTeD current development to design and develop activities within text-mining projects for microbiology community.

**Keywords:** user need analysis, text-mining, microbial diversity

## 1 Introduction

The microbial diversity field encompasses all questions related to the variety and the variability of microorganisms and their ecosystems. It is then related to most activities in the microbiology field, which include human, plant or animal health, but also to positive effects of microorganisms, such as plant growth, bioremediation or food processing. As in every Life Sciences domain, microbial diversity information is spread among many complementary sources of experimental and curated data and among huge amounts of textual documents, such as scientific publications, industrial technology reports or medical documents. Text-mining (TDM) and information extraction has been recognized as an efficient way to extract and formalize information in Life Sciences so that it can then be combined with other sources of data. Compared to other Life Sciences domains such as molecular mechanism study which have been largely targeted

by text-mining research, the microbial diversity domain suffers from a lack of attention. The increasing need of combining genetics information with ecology information motivates the development of dedicated TDM solutions for microbial diversity studies [4]. This justifies its choice as a use case domain for the European TDM infrastructure *OpenMinTeD* (OMTD). A first critical step is the analysis of the needs of the microbiology community with respect to diversity questions.

End-users like biologists rarely interact directly with the TDM tools but rather consume the results through third party applications. Starting from their needs helps identifying the other stakeholders - from bioinformaticians to legal experts - who will build those TDM based solutions for them.

### 1.1 Biological Context

Microorganisms are defined as microscopic living organisms. They are very abundant, even in places where conditions are extreme (*e.g.* with high and low pressure, high or low temperatures). Any physical location, from the smallest (*e.g.* microscopic) to the largest (*e.g.* a planet) is a habitat for microorganisms. Microbial diversity research aims to study and describe microbiomes, to understand microorganism interactions, their ecosystems, how they adapt to their ecosystem [1] and their phylogeny. The recent advances in high throughput molecular technologies, such as DNA-sequencing and metagenomics have deeply changed biodiversity research methods. Indeed, biology and bioinformatics projects produce huge amounts of heterogenous information about the genetic sequences and the species and strains that have been experimentally identified in a given environment [8]. The availability of tools for the joint analysis of biotopes, taxa and genetic sequences is an emerging need which is the focus of this work. The information on organism biotopes is a crucial knowledge in biodiversity research. The quantity of microbial biotope descriptions increases with the need for fast and high-scale comparison among microorganism ways of life. While this is a significant community need, there is no resource that centralizes and formalizes the knowledge on microorganism habitats, which makes it hard to look up and compare information. This motivated the development of habitat classifications as illustrated by the call for standardized classification of metagenome projects [10] or the OntoBiotope ontology [3] development.

### 1.2 Microbiology and Text and Data Mining

Database information and literature information are complementary. However free text fields in databases as well as scientific papers are expressed in natural language. The formalization of the descriptions by Text and Data Mining (TDM) [15] is then a compulsory step for any further processing because of the high variability of the text. The resulting formal representation should allow for

[3] Available on *Agroportal* website

the automatic analysis of descriptions of microorganism biotopes from different experiments, which could then be compared at a large scale.

Information content analysis and standardization requires TDM tools to automatically extract relevant spans of text, and normalize or categorize them with reference resources such as ontologies.

### 1.3 User analysis context

This requirement analysis has been brought about by the OpenMinTeD (OMTD) project. The OMTD project aims to build a text-mining infrastructure for both application domain users and text-mining experts. Guided by the needs of the microbiologists who are the targeted end-users of our TDM solutions, content providers, aggregators, and application developers find themselves in the position to seek out TDM solutions that can be incorporated into their services. Bioinformatics infrastructures play an important role in the microbiology domain.They cover a wide range of activities, including information and data management and processing. These infrastructures aggregate data and build and operate the applications used by the researchers. As such they are also targeted by the requirement analysis in order to propose adequate tools and methods. The content and resource providers, legal experts and socio-economic actors are also considered here as strategic stakeholders for the adoption and success of TDM solutions, and so, the user analysis was also performed for them [4]. A stakeholder map helps to understand and to validate the feasibility of a microbial end-user application.

Beyond the OMTD project, this study has been conducted with the general purpose of providing the TDM community with an inventory of the microbiology community expectations that take into account the recent evolution of both the microbiology and the TDM domains.

## 2 Approach

The OMTD project targets four different research community domains, *Scholarly Communication*, *Life Sciences*, *Social Sciences* and *Agriculture and Biodiversity* through different example of use-case. This work takes place in the *Agriculture and Biodiversity* field. Needs analysis was carried out from September, 2015 to March, 2016 using various methods: interviews, a plenary session, workshops and meetings with other TDM partners. The methodology is fully described in the *Requirement Methodology* report [5]. It follows three phases, briefly sketched here:

[4] For more details on the user analysis for all stakeholders, please read the document *White paper on OpenMinTed Community Requirements* available on the *OpenMinTeD* website

[5] It is available on the *OpenMinTeD* website

- Phase 1 (the preliminary phase) identified end-users and a first set of their needs in terms of TDM services. This phase determined the main scope of the use-case with respect to the expectations of the end-users. It recorded a preliminary outlook of the end-users towards issues in relation to TDM such as access to content, licensing or storage.
- Phase 2 identified the main stakeholders (data provider, producer and consumer) in this use-case, and their interactions. This analysis has been shared across all OMTD application domains.
- Phase 3 validated the requirements through final interviews with end-users. Exchanges and discussions with the OMTD partners also allowed the identification of the final main stakeholder types, which are generalizable to all TDM use-cases. In parallel, a list of requirements was established to guide the functional specifications of TDM solutions.

## 3 Results of the user needs analysis

### 3.1 Preliminary needs and scope of the study

The first phase allowed us to identify the potential end-user stakeholders and the scope of our TDM application with a peculiar focus on the food domain, particularly relevant for our research institute. The potential end-users which identified are numerous and diverse: research laboratories, agrofood companies, food safety agencies, agro-industrial technology institutes and research network. Indeed, these different profiles have different aims. In order to better understand them, we describe here their typology:

*Research laboratories* The microorganisms specifically added during the food process are usually known, but the environmental microorganisms that could interact with the artificial microbiota and modify its composition, are mostly unknown. Microorganisms may contribute to desirable properties or defects of the final food product (due to their phenotype and the growth context). Fundamental microbiology research needs a better understanding of the role of microorganisms, whether pathogenic or with beneficial effects for health.

*Agro-industrial Technology Institutes* One role of the technology institutes is to foster the link between fundamental research and companies, to provide access to research and development results. By dissemination of knowledge and technology innovation, these institutes provide diagnosis, counseling and supports.

*Agrofood Companies* Research in food companies pursues technical innovation such as the development of new food products with increasing economic impact. Industrial important characteristics of the microbial flora are food safety, organoleptic and nutritional benefits. There is a focus on positive flora: for instance, to produce fermented foods (*i.e.* yoghurt, cheese), companies use starter cultures with known microorganism flora added as ingredients. Bacteria also produce substances that are used for the design of new food products [12].

*Food Safety Agencies* Food Safety Agencies tasks are health risk monitoring, expertise, research and supply of reference data on food risk. The evaluation

of the risk of foodborne illness is done through the information feedback from analytical laboratories. Tracing back pathogenic bacteria biotopes would help identify their origin (food ingredient, production process or transport) and control their occurrence.

*Research Networks* National and international research networks disseminate information among partners and coordinate information sharing and publication efforts. A focus is on microbiome description standard, and specifically on genetics and biodiversity.

A meeting with representatives of these targeted communities (about twenty participants) defined the framework for our study. All the participants, either researchers, companies or service providers, advocate a formal and unified representation of microorganism biotope information from various sources as a first critical step.

At the end of this preliminary study, we focused on a use-case targeting global microbial biodiversity expressed in texts written in English. This first phase confirmed also the interest in food products, and in particular in dairy products. The microbial scope of the study centered on any type of microorganism considered as positive flora (thus excluding pathogens). These first potential end-user communities may be included in or excluded from the final use-case, based on the general expected outcomes.

### 3.2 Stakeholders of TDM solutions applied to Microbiodiversity

The second phase is dedicated to the description of all the stakeholders involved in the target TDM application beyond the end-users, as well as their interactions. Indeed, preliminary discussions with end-users identified different types of stakeholders of the TDM solution. Then, focus group meetings, workshops and interviews with forty persons enriched the initial need description performed at the microbiologist community level in the first phase. These meetings and interviews were held with representatives of the identified stakeholders, within and outside the project consortium: microbiology researchers, risk analyzers, database developers and infrastructure managers. We classified them into five main types: (1) the end-users, including researchers and industrial communities (2) e-infra operators and aggregators, including application developers and knowledge managers; (3) text-miners and then (4) content providers and (5) curators, whose role is essential to improve the final results of TDM extraction. The stakeholders and their interactions are summarized in figure 1. The main characteristics of these groups are briefly detailed below:

- End-Users are characterized as the text mining service consumers; they may have a socio-economic interest.
- Curators are domain specialists (from research scientific or industrial communities), who ensure high quality of data. Their role is the cleaning of the text-mining results and the revision of the domain specific resources to improve the final quality.

– Data and services aggregators - infrastructure providers want to extend the scope of the services they offer.
– TDM experts, either researchers or engineers, develop tools and services to be used by the TDM infrastructure
– Content providers provide textual and semantic resources that are processed by the TDM infrastructure

Stakeholders we identified appear to be common to other use-cases from the *Life Sciences* and *Agriculture and Biodiversity* domains, where e-infrastructure and data aggregators also play an important role as in all experimental domains. We also identified legal experts as general stakeholders, that share expertise on legal aspects such as international copyright or intellectual property laws for design of TDM solutions. For this work on the microbial diversity domain, we focus our need and requirement analysis on the end-user group, which expressed specific needs (red rectangle in figure 1). Needs of other stakeholders are less specific to this use case, are not be detailed here, but help us to take a step back from the microbiologist needs.

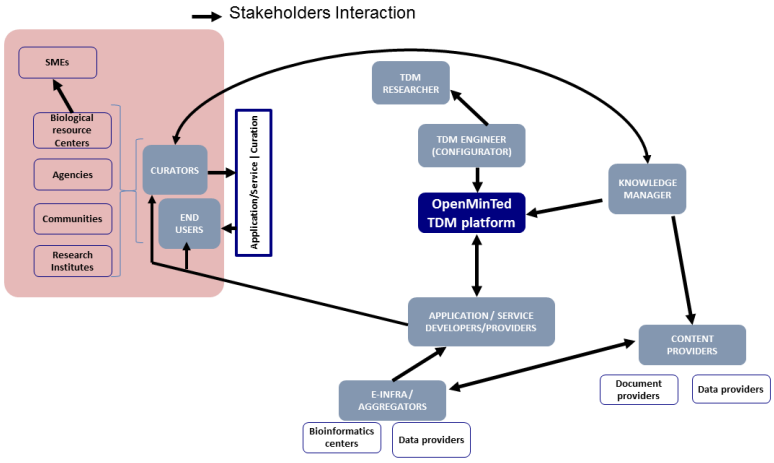


**Fig. 1.** Interactions among stakeholders of the Microbial Biodiversity Use Case

### 3.3 User-centered method for end-users need analysis

Phase 1 targets the sub-community of the positive flora in the food microbiology domain, which involves all types of end-users, including companies. User need analysis may be difficult to conduct in domains where user profiles, practices, experienced difficulties and expectations are very diverse, which is the case in the

microbial diversity domain. In order to deal with this situation and to express requirements of end-users, we used a user-centered method.

The user-centered methodology revolves around profiling and analyzing users, in order to identify the specific characteristics of different types of users. This was carried out through the persona methodology [7]. Personas are not real persons but fictional representations of a user, which could represent real people during the design process.

The aim is to determine which type of integrated information and which application interface fit the needs of each persona. We further want to detail and prioritize aims and behaviors of personas regarding TDM solutions. To do this, we used an approach comparable to Calde et al. [5] to design our personas. Phase 1 and 2 allowed us to roughly sketch the profile of some representative personas, and we selected four of them. To complete the description of personas, we organized a dedicated one-day workshop that involved French microbiologists from Inra, the French National Institute for Agricultural Research. Participants were split into four groups, one group per persona. In order to guide the formulation of the description, we provided four questions to be answered by the participants, adapted from the Lean Canvas [14]: (1) Who is the persona? (name, age, function, skill) (2) What kind of information is s/he interested in? (3) What kind of access/needs would be useful? and (4) What solutions would satisfy his/her needs and those of his/her community? The persona description is provided in Table 1. The participants in the different groups also designed collaboratively the wireframes of the application corresponding to each personas needs. Figure 2 presents an example of the access to the TDM results for persona 3 ("Patrice"). It enables the user to search strains associated to a phenotype extracted by text-mining, and provides the link to the original publications so that the user can check the experimental setting and the relevance of the result. The results on persona and application design were then shared among all groups through presentations at the end of the day.

Based on the work of Calde et al.[5], we defined our personas as "secondary personas", that is, a main application interface will serve the needs of the different types of secondary persona with minor modifications/additions .

### 3.4 From persona to end-users needs

Personas allowed us to identify various specific needs with respect to the food ecosystem knowledge. We refer to them using their fictional first names to illustrate our interpretation (see Table 1 for the detailed needs of persona).

*Microbial biodiversity* All personas expressed a need related to knowledge on microorganism identification and provenance, *i.e.* to study and characterize the microbial biodiversity of food ecosystems (such as Yann and Claudia), and to compare food biodiversity to other ecosystems in order to qualify microorganism provenance (such as Lily and Patrice). Moreover, from the point of view of traceability, knowledge about the known living places of microorganisms would contribute to prevent contamination, either in the medical environment or in the food industry (as expressed by Yann). An extension of this need is a tool that

**Table 1.** Persona descriptions

| | |
|---|---|
| **Yann, 35 years old, SME company in bio-preserved food product** | |
| What information? | Information and literature about bio-preservation process to answer the following questions: what non-pathogenic bacteria grows in a given type of food and contributes to longer preserve the food quality? What are the chemical compounds produced or degraded by the bacteria that might interfere with given food products? |
| What needs? | An easily accessible database with a simple and user-friendly search engine, that he can query about the biopreservation of the type of food products he is interested in. |
| What solutions? | A database of the bacterial strains that contribute to biopreservation linked to the food products of interest and that exhibit the relevant growth phenotype (*e.g.* pH, temperature...). The database should include information about the public or private institutes that provide the strain, the patent if any, as well as organoleptic properties and health impact. |
| **Lily, 38 years old, former baker and expert in bread leavens** | |
| What information? | She wants to compare the functional diversity of commercial and artisanal leavens. She uses the information from scientific papers, collections of national and international organisms, local expertise and expertise from baker networks |
| What needs? | She needs to know about the habitats of microorganisms used in bread making, the physico-chemical characteristics of starter cultures for bread and the phenotypes of the species listed in leavens. |
| What solutions? | Pooling data obtained on bacteria and yeasts (and their interaction), with an analysis of phenotypes (at strain, species and leaven levels). Using the knowledge about the taxonomic diversity of yeasts together with phenotype information of strains, species and leaven in order to compare the functional diversity of commercial and artisanal leavens. |
| **Claudia, 40 years old, bacteria physiology researcher expert in aroma compounds produced by lactic bacteria** | |
| What information? | She wants to use new strains for the production of aroma in new food products. She first needs information about aroma compounds that bacteria produce and about where these bacteria live. She collects samples of cheese or dairy products, and she traces the technological process of these products: which milk? Which treatment? Which manufacturing process? She uses metagenomics experiments (DNA and RNA information such as RNA 16s) to identify bacteria strains. |
| What needs? | She would need access to a database on the bacterial metagenomic and phenotypic data, with non-redundant and non-contradictory information. She works on little known species, so she needs a reliable database to identify and compare their own genetic data to those of the database, and identify where they come from. |
| What solutions? | Pooling of data from text in a database about bacteria genotypes and phenotypes that she could query with a bacterial phenotype using a user-friendly interface. |

*continued on next page*

*continued from previous page*

| **Patrice, 55 years old, microbiologist engineer working in the R&D department of a dairy company** | |
|---|---|
| What information? | He wants to gather information for the design of new plant food products obtained by fermentation. He uses scientific literature (PubMed bibliographic database, articles, journals, and conference reports) and student reports and he interacts with technical centers and researchers. |
| What needs? | He needs a tool that allows him to quickly identify and get a bacterial strain that meets his requirements: for example, milk and cereal growth medium, degradation of sugar lactose, production of the buttery flavor diacetyl molecule. |
| What solutions? | A database that associates strains, phenotypes and growth conditions (*e.g.* lactose consumption, diacetyl production acidification capacity), and the link to publication for more information. |

could be used in the food risk analysis community, enabling cross-checking in an easy manner by safety agencies.

*Microbial phenotypes* A second set of needs is related to knowledge on microorganisms. Microorganisms are massively used in food industry, firstly as a fermentative agent (as expressed by Lily and Patrice) but not only. Their peculiarity is the production of some nutritional or sensory compounds of interest, such as aromas (as illustrated by Claudia and Yann). Furthermore, a given microorganism can either have a positive or an undesirable effect, according to the context in which it produces the molecule. The study of molecules produced by bacteria gathers a large community of researchers [13].

*Industrial uses* There is a need to study and characterize the phenotypes of microorganisms for food quality improvement, *e.g.* flavor and taste, nutriments, biopreservation or optimization of food engineering processing. Industrial researchers constantly look for new food products to meet specific diets and special dietary needs (such as Yann and Patrice). A database referencing knowledge on microbial phenotypes would allow researchers to easily detect a species that is able to degrade or consume a specific molecule. Biologists want to know which molecule was produced by which strain, but also in which substrate, food or habitat the strain can live. The challenge for biologists is to link bacteria, habitats including artificial media and produced molecules. Without the information on growth media, the production of the molecule by the bacteria may be jeopardized. This information is generally expressed in scientific publications, magazines, tables, or personal databases (as described by Patrice). In the specific case of new food processing, the integration of heterogenous data would enable users to quickly search for useful microorganisms (data extracted from the literature and existing genomics databases such as detailed by Claudia).

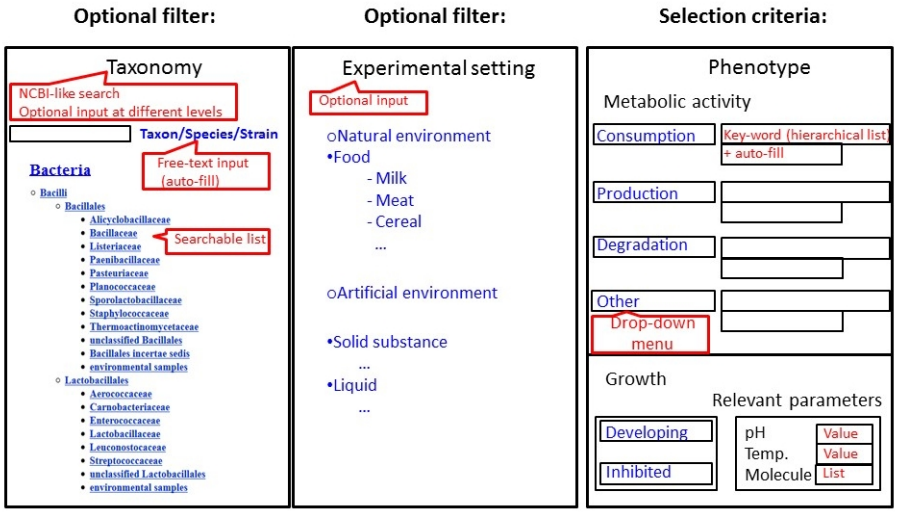


**Fig. 2.** Application wireframe for the persona "Patrice", the microbiologist engineer

## 4 Scenario: from end-users analysis to TDM solutions

Based on this analysis, we designed a microbial use-case scenario to assist microbiology sub-communities research by TDM. Three main needs were identified, that is the completion of the current microbial knowledge with respect to their biodiversity, phenotypes and usability. The scope of the TDM application was defined as the recognition and normalization of relevant entities, namely microorganisms, habitats, molecules, applications and phenotypes, in scientific publications; and the identification of the relationships between these entities. One specific final application will focus on positive flora, and in particular on dairy food. To do this, we aim to leverage state-of-the-art tools and resources. However, until recently, there have been few attempts to apply text-mining to microorganism biodiversity study. The Bacteria Biotope task of the BioNLP Shared Task in 2011 [4], 2013 [3] and 2016 [9] aimed at the extraction and normalization of bacteria and habitats from text, and their linking. Dedicated text-mining methods, benchmarks to evaluate these methods, and resources were developed and will be reused for our use-case. A main need is the normalization of TDM results with knowledge resources (*e.g.*, taxonomies, ontologies) so that they can be aggregated, integrated and compared according to a same reference. There are few standard resources for the food domain with respect to microbial point of view. The *OntoBiotope ontology* was built for the description and the normalization of microorganism habitats. Another existing ontology is the Environment Ontology (*EnvO*), even if the description level appears more relevant for macroscopic habitats. Indeed, it has been shown that OntoBiotope yields better results than the general purpose EnvO for microbiology habitats [16] [6]. The EFSA classification *FoodEx2* part on food contamination is also a valuable

source. However, these specific resources may need adaptation to fully meet the needs of microbiologist users.

The automatic extraction of molecules and microbial phenotype descriptions is a new open research question in the food domain. Previous results obtained on molecule name extraction (*e.g.* BioCreative IV CHEMDNER Task) will be reused [11]. The expressions of phenotypes at various levels (molecular, physiological and application) are more difficult to handle and may require the development of a new ontology.

Finally, the TDM scenario includes the extraction of relations between normalized entities, such as *a given microorganism produces a given molecule*. Supervised machine learning methods trained with manually annotated examples will be used. Generic methods such as TEES [2] or AlvisRE [17] have demonstrated their efficiency when applied to relation extraction in biology. The TDM results will then be aggregated with other heterogenous data in the bioinformatics application, and will be available to the biological end-users.

## 5 Conclusion

The elicitation of TDM requirements in the food microbiology domain confirms that there is a mosaic of needs. We identified the main needs from fictional users and representatives of the sub-community of the food positive flora microbiologists, that are the targeted end-users. TDM users want to study, characterize and compare the microbial biodiversity of food ecosystems. Another need is the discovery of information among large sets of articles. Aggregation of heterogenous data and relevant information from various sources is also a strong requirement. The persona based approach allowed us to design precise scenarii involving complex TDM workflows, adequate existing knowledge resources, and relevant textual sources. To answer the end-user needs, we involved intermediate stakeholders who are currently working on the integration of the TDM results in their bioinformatics applications. These detailed user needs also allowed us to participate in the definition of the OpenMinTeD platform concrete requirements that ground its implementation.

## Acknowledgements

This work was supported by the French National Institute for Agricultural Research (Inra) and the OpenMinTeD project (EC/H2020-EINFRA 654021). This work was carried out with the technical help of Patricia Gerretto (Inra) and Vivi Katifori (Agroknow). We would like to thank biologists, bioinformaticians and all persons who have participated in the end-user need analysis; and in particular the Inra Florilège working group, Food Microbiome project, Biological Resource Centers, Bioinformatics infrastructure Migale from IFB (French Institute of Bioinformatics), ANSES (the French Agency for Food, Environmental and Occupational Health & Safety), JGI (Joint Genome Institute) and GBIF (Global Information Facility).

## References


1. Bent, S.J., Forney, L.J.: The tragedy of the uncommon: understanding limitations in the analysis of microbial diversity. The ISME journal 2(7), 689–695 (2008)
2. Björne, J., Salakoski, T.: Tees 2.1: Automated annotation scheme learning in the bionlp 2013 shared task. In: Proceedings of the BioNLP Shared Task 2013 Workshop. pp. 16–25. Association for Computational Linguistics (2013)
3. Bossy, R., Golik, W., Ratkovic, Z., Valsamou, D., Bessières, P., Nédellec, C.: Overview of the gene regulation network and the bacteria biotope tasks in bionlp'13 shared task. BMC bioinformatics 16(10), S1 (2015)
4. Bossy, R., Jourde, J., Manine, A.P., Veber, P., Alphonse, E., Van De Guchte, M., Bessières, P., Nédellec, C.: Bionlp shared task-the bacteria track. BMC bioinformatics 13(11), S3 (2012)
5. Calde, S., Goodwin, K., Reimann, R.: SHS Orcas: The first integrated information system for long-term healthcare facility management. In: Case Studies of the CHI2002. pp. 2–16. ACM (2002)
6. Cook, H.V., Pafilis, E., Jensen, L.J.: A dictionary-and rule-based system for identification of bacteria and habitats in text. ACL 2016 p. 50 (2016)
7. Cooper, A., Reimann, R., Cronin, D., Noessel, C.: About face: the essentials of interaction design. John Wiley & Sons (2014)
8. Dahllöf, I.: Molecular community analysis of microbial diversity. Current Opinion in Biotechnology 13(3), 213 – 217 (2002)
9. Deléger, L., Bossy, R., Chaix, E., BA, M., Ferré, A., Bessières, P., Nédellec, C.: Overview of the bacteria biotope task at BioNLP shared task. In: BioNLP Shared Task. p. 101. The Association for Computational Linguistics, Berlin, Germany (Aug 2016)
10. Ivanova, N., Tringe, S.G., Liolios, K., Liu, W.T., Morrison, N., Hugenholtz, P., Kyrpides, N.C.: A call for standardized classification of metagenome projects. Environmental microbiology 12(7), 1803–1805 (2010)
11. Krallinger, M., Leitner, F., Rabal, O., Vazquez, M., Oyarzabal, J., Valencia, A.: Chemdner: The drugs and chemical names extraction challenge. Journal of cheminformatics 7(1), S1 (2015)
12. Leroy, F., De Vuyst, L.: Fermented food in the context of a healthy diet: how to produce novel functional foods? Current Opinion in Clinical Nutrition & Metabolic Care 17(6), 574–581 (2014)
13. Longo, M.A., Sanromán, M.A.: Production of food aroma compounds: microbial and enzymatic methodologies. Food Technology and Biotechnology 44(3), 335–353 (2006)
14. Maurya, A.: Running lean: iterate from plan A to a plan that works. O'Reilly Media, Inc. (2012)
15. Pignatelli, M., Moya, A., Tamames, J.: Envdb, a database for describing the environmental distribution of prokaryotic taxa. Environmental Microbiology Reports 1(3), 191–197 (2009)
16. Ratkovic, Z., Golik, W., Warnier, P.: Event extraction of bacteria biotopes: a knowledge-intensive nlp-based approach. BMC bioinformatics 13(11), S8 (2012)
17. Valsamou, D.: Information extraction from scientific articles for the reconstruction of the biological regulatory network during the seed development phase of *Arabidopsis thaliana*. Ph.D. thesis, Université Paris-Saclay, France (2017)