Automatic data extraction from EFSA Opinions

Diana Varšíková, Ecomole s.r.o

August 27, 2024

1 Introduction

We received csv file with 204 question numbers/opinions to include in the database. Below is an example of the data that are to be extracted:

Administrative Data of the Opinion: Applicant company, their country of origin, DOI of the opinion, plus 10 other fields.
General Information: Trade names, common names, form of the food (whole food, extract, etc.).
Identity: This depends on the category. For example, for animals: genus, species, subspecies, part used; for chemicals: common name, IUPAC name, CAS number, molecular formula, etc.
Production Process: List of main production steps.
Main Composition of the Novel Food: Carbohydrates, proteins, fats, minerals, water, vitamins.
Proposed Uses and Use Levels: Proposed uses (whole foods, supplements, etc.), target population (general, infants, children, etc.), food category (FoodEx / FAIM).
Nutritional Information: Nutritionally disadvantageous / nutritionally advantageous components.
Availability of ADME (Absorption, Distribution, Metabolism, and Excretion) Studies.
Availability of Toxicological Studies and Their Outcome.
Allergenicity Assessment: Unlikely, low, possible, certainty etc.

The primary objective was to automate the extraction of 'non-expert' information—data that can be easily understood and assessed for accuracy by individuals without a biology degree. This approach allows experts to focus on more complex and technical information. The goal was to automatically extract the majority of the administrative data and some elements of the general information This report will detail which data were successfully extracted automatically and the logic and methods used in the extraction process.

2 Data Summary

We received a CSV file from EFSA containing a list of 207 question numbers to be included in the database. The types of articles associated with these question numbers are as follows:

Type Of Article	Count
Opinion	198
Guidance	4
Public consultation	3
Data gaps statement	1

This distribution indicates that 198 opinions are to be extracted, as it was agreed not to extract data from the other types of articles. Out of the 198 opinions, EFSA provided Ecomole with JATS files for 194 of them. However, there are variations in the completeness of these files:

JATS availability	count
full JATS file	123
front JATS file	71
no JATS file	4

This summary highlights that 71 of the JATS files contain only the FRONT section, which includes administrative information but lacks the content part necessary for a full extraction.

3 Opinions analysis

By looking at the opinions, the structure of the opinions can be divided into five groups.

Group A

This is the most common group.

SCIENTIFIC OPINION

EFSA Journal

ADOPTED: 13 December 2016

doi: 10.2903/j.efsa.2017.4682

Scientific Opinion on taxifolin-rich extract from Dahurian Larch (*Larix gmelinii*)

EFSA Panel on Dietetic Products, Nutrition and Allergies (NDA),
Dominique Turck, Jean-Louis Bresson, Barbara Burlingame, Tara Dean, Susan Fairweather-Tait, Marina Heinonen, Karen Ildico Hirsch-Ernst, Inge Mangelsdorf, Harry J. McArdle, Androniki Naska, Monika Neuhäuser-Berthold, Grazyna Nowicka, Kristina Pentieva, Yolanda Sanz, Alfonso Siani, Anders Sjödin, Martin Stern, Daniel Tomé, Marco Vinceti, Peter Willatts, Karl-Heinz Engel, Rosangela Marchelli, Annette Pöting, Morten Poulsen, Josef Schlatter, Wolfgang Gelbmann and Henk Van Loveren

Abstract

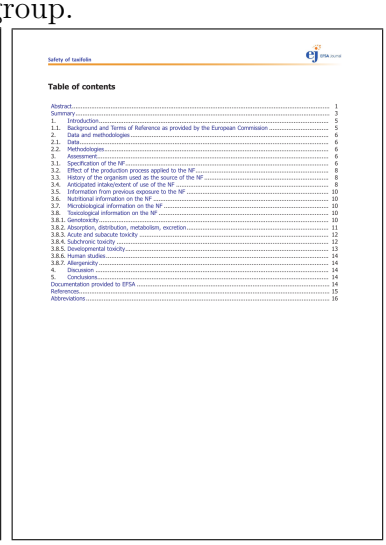
Following a request from the European Commission, the EFSA Panel on Dietetic Products, Nutrition and Allergies (NDA) was asked to carry out the additional assessment for taxifolin-rich extract from Dahurian Larch as a food ingredient in the context of Regulation (EC) No 258/97. The novel food (NF) is a taxifolin-rich water -ethanol extract from the wood of the Dahurian Larch and contains a minimum of 90% taxifolin. The Panel considers that the taxifolin-rich extract is sufficiently characterised and that its compositional data and specifications do not raise safety concerns. The NF is intended to be added to non-alcoholic beverages, to yogurt and to chocolate confectionery. The Panel considers that the data on genotoxicity do not raise concern. In a subchronic rat study performed in accordance with OECD standards, the highest dose tested (i.e. 1,500 mg/kg bw) was considered to be the NOAEL. The margin of exposure (MOE) of the combined intake (158 mg) from the intended food uses (including 100 mg from food supplements) would result to about 660 for an adult weighing 70 kg. For adolescents, taking into account a default body weight of 45 kg, the MOE of the combined intake (146 mg) would be about 460. In the absence of a high percentile intake estimate for children between 9 and 14 years of age, the Panel considers the P97.5 intake estimate from the intended food uses (except from food supplements) for children between 10 and 17 years, i.e. 46 mg/day. Taking into account a default body weight of 29.4 kg (P5 body weight for children aged 10-14 years as suggested by EFSA Scientific Committee (2012)), the resulting MOE would be about 960.

Keywords: taxifolin, (2*R*,3*R*) *trans*-dihydroquercetin, 2,3-dihydroquercetin, Dahurian Larch, novel food, ingredient

Requestor: European Commission following an application by Ametis JSC

Question number: EFSA-Q-2012-00961

Correspondence: nda@efsa.europa.eu

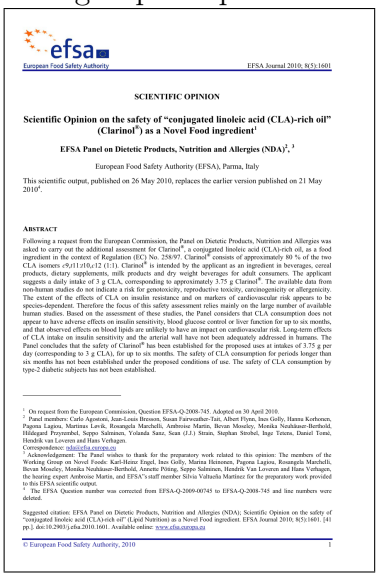


Safety of taxifolin

Table of contents

Group B

This group has panel in the footnote and abstract on the first page.



efsa
European Food Safety Authority
EFSA Journal 2010; 8(5):1601

SCIENTIFIC OPINION

Scientific Opinion on the safety of "conjugated linoleic acid (CLA)-rich oil" (Clarinol®) as a Novel Food ingredient[1]

EFSA Panel on Dietetic Products, Nutrition and Allergies (NDA)[2, 3]

European Food Safety Authority (EFSA), Parma, Italy

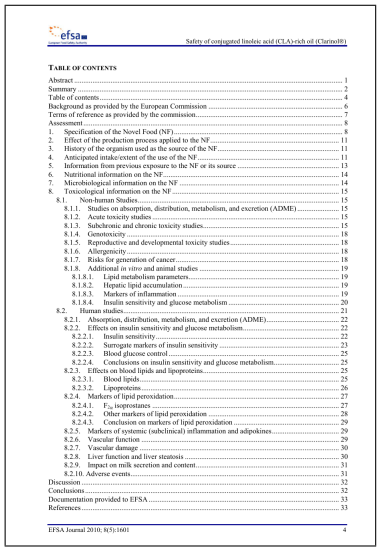
This scientific output, published on 26 May 2010, replaces the earlier version published on 21 May 2010[4].

ABSTRACT

Following a request from the European Commission, the Panel on Dietetic Products, Nutrition and Allergies was asked to carry out the additional assessment for Clarinol®, a conjugated linoleic acid (CLA)-rich oil, as a food ingredient in the context of Regulation (EC) No. 258/97. Clarinol® consists of approximately 80 % of the two CLA isomers *c*9,*t*11-c*t*10,*c*12 (1:1). Clarinol® is intended by the applicant as an ingredient in beverages, cereal products, dietary supplements, milk products and dry weight beverages for adult consumers. The applicant suggests a daily intake of 3 g CLA, corresponding to approximately 3.75 g Clarinol®. The available data from non-human studies do not indicate a risk for genotoxicity, reproductive toxicity, carcinogenicity or allergenicity. The extent of the effects of CLA on insulin resistance and on markers of cardiovascular risk appears to be species-dependent. Therefore the focus of this safety assessment relies mainly on the large number of available human studies. Based on the assessment of these studies, the Panel considers that CLA consumption does not appear to have adverse effects on insulin sensitivity, blood glucose control or liver function for up to six months, and that observed effects on blood lipids are unlikely to have an impact on cardiovascular risk. Long-term effects of CLA intake on insulin sensitivity and the arterial wall have not been adequately addressed in humans. The Panel concludes that the safety of Clarinol® has been established for the proposed uses at intakes of 3.75 g per day (corresponding to 3 g CLA), for up to six months. The safety of CLA consumption for periods longer than six months has not been established under the proposed conditions of use. The safety of CLA consumption by type-2 diabetic subjects has not been established.

[1] On request from the European Commission, Question EFSA-Q-2008-745. Adopted on 30 April 2010.
[2] Panel members: Carlo Agostoni, Jean-Louis Bresson, Susan Fairweather-Tait, Albert Flynn, Ines Golly, Hannu Korhonen, Pagona Lagiou, Martinus Løvik, Rosangela Marchelli, Ambroise Martin, Bevan Moseley, Monika Neuhäuser-Berthold, Hildegard Przyrembel, Seppo Salminen, Yolanda Sanz, Sean (J.J.) Strain, Stephan Strobel, Inge Tetens, Daniel Tomé, Hendrik van Loveren and Hans Verhagen.
Correspondence: ndatotf@efsa.europa.eu
[3] Acknowledgement: The Panel wishes to thank for the preparatory work related to this opinion: The members of the Working Group on Novel Foods: Karl-Heinz Engel, Ines Golly, Marina Heinonen, Pagona Lagiou, Rosangela Marchelli, Bevan Moseley, Monika Neuhäuser-Berthold, Annette Pöting, Seppo Salminen, Hendrik Van Loveren and Hans Verhagen, the hearing expert Ambroise Martin, and EFSA's staff member Silvia Valtuena Martinez for the preparatory work provided to this EFSA scientific output.
[4] The EFSA Question number was corrected from EFSA-Q-2009-00745 to EFSA-Q-2008-745 and line numbers were deleted.

Suggested citation: EFSA Panel on Dietetic Products, Nutrition and Allergies (NDA); Scientific Opinion on the safety of "conjugated linoleic acid (CLA)-rich oil" (Lipid Nutrition) as a Novel Food ingredient. EFSA Journal 2010; 8(5):1601. [41 pp.]. doi:10.2903/j.efsa.2010.1601. Available online: www.efsa.europa.eu



efsa
Safety of conjugated linoleic acid (CLA)-rich oil (Clarinol®)

TABLE OF CONTENTS

Group C

This group has summary and keywords on the first page, and panel members on the last page.

efsa
European Food Safety Authority
The EFSA Journal (2007) 580, 1-10

Opinion of the Scientific Panel on Dietetic Products, Nutrition and Allergies on a request from the Commission related to the safety of *Allanblackia* seed oil for use in yellow fat and cream based spreads

(Request N° EFSA-Q-2007-059)

(Adopted on 25 October 2007 by written procedure)

SUMMARY

Allanblackia seed oil is obtained from the seeds of the *Allanblackia* tree (*A. floribunda* and *A. stuhlmannii*). The oil is isolated and refined according to commonly applied procedures. It meets the quality criteria of edible vegetable oils and the limits for potential contaminants set in the EU. Its physicochemical properties (semi-solid at room temperature) are determined by the high contents of stearic-oleic-stearic (SOS) and stearic-oleic-oleic (SOO) triglycerides (average values 69 % SOS; 23.0 % SOO). This unique arrangement of fatty acids makes the oil suitable as a hardstock component. Accordingly, *Allanblackia* seed oil is intended for use as a novel food ingredient in yellow fat and cream-based spreads.

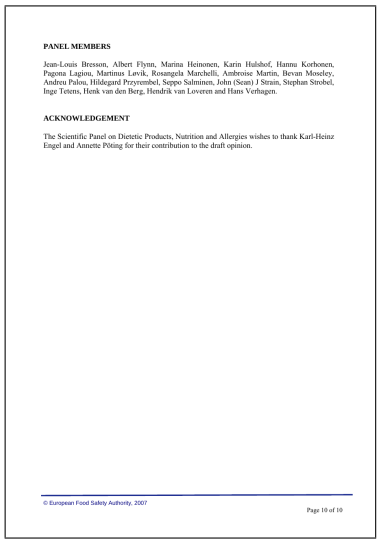
The intake of *Allanblackia* seed oil was estimated on the basis of margarine and yellow fat spread consumption data from Germany, Sweden, the United Kingdom and the Netherlands, assuming that total margarine and yellow fat spread consumption will be replaced by products containing 20 % *Allanblackia* seed oil. Based on a 95th percentile margarine intake of 39 g/person/day across all age-groups in Germany, an intake of approximately 8 g of *Allanblackia* seed oil/person/day was estimated. On the basis of consumption data for yellow fat spreads in Sweden, the UK and The Netherlands, the applicant estimated *Allanblackia* seed oil intakes (95th percentile) to range from 2.8 g /person/day, for British boys and girls aged from 1.5 to 3.5 years, to 16 g/person/day, for Swedish women aged from 18–74 years.

Genotoxicity tests revealed no evidence of genotoxic/mutagenic properties. In a 90-day feeding study with rats, the administration of 20 % *Allanblackia* seed oil did not induce any toxicologically relevant effects which were attributable to administration of *Allanblackia* seed oil and which are not also seen in animal studies with other high fat diets.

The Panel concludes that *Allanblackia* seed oil is safe for human consumption under the specified conditions of use.

KEY WORDS

Allanblackia, seed oil, yellow fat spreads, cream-based spreads, novel food ingredient



PANEL MEMBERS

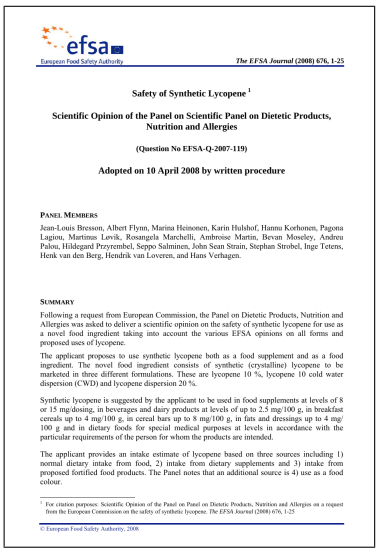
Jean-Louis Bresson, Albert Flynn, Marina Heinonen, Karin Hulshof, Hannu Korhonen, Pagona Lagiou, Martinus Løvik, Rosangela Marchelli, Ambroise Martin, Bevan Moseley, Andreu Palou, Hildegard Przyrembel, Seppo Salminen, John (Sean) J Strain, Stephan Strobel, Inge Tetens, Henk van den Berg, Hendrik van Loveren and Hans Verhagen.

ACKNOWLEDGEMENT

The Scientific Panel on Dietetic Products, Nutrition and Allergies wishes to thank Karl-Heinz Engel and Annette Pöting for their contribution to the draft opinion.

Group D

This group has panel members on the first page, followed by summary and keywords.



Safety of Synthetic Lycopene [1]

Scientific Opinion of the Panel on Scientific Panel on Dietetic Products, Nutrition and Allergies

(Question No EFSA-Q-2007-119)

Adopted on 10 April 2008 by written procedure

PANEL MEMBERS

Jean-Louis Bresson, Albert Flynn, Marina Heinonen, Karin Hulshof, Hannu Korhonen, Pagona Lagiou, Martinus Løvik, Rosangela Marchelli, Ambroise Martin, Bevan Moseley, Andreu Palou, Hildegard Przyrembel, Seppo Salminen, John Sean Strain, Stephan Strobel, Inge Tetens, Henk van den Berg, Hendrik van Loveren, and Hans Verhagen.

SUMMARY

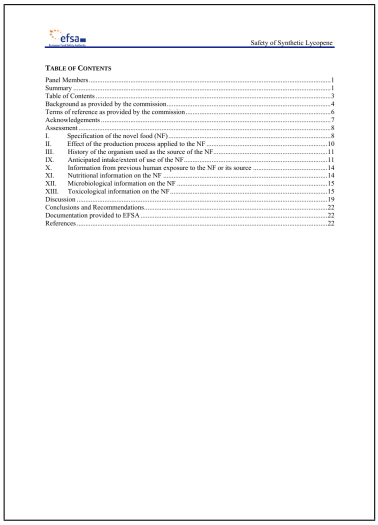
Following a request from European Commission, the Panel on Dietetic Products, Nutrition and Allergies was asked to deliver a scientific opinion on the safety of synthetic lycopene for use as a novel food ingredient taking into account the various EFSA opinions on all forms and proposed uses of lycopene.

The applicant proposes to use synthetic lycopene both as a food supplement and as a food ingredient. The novel food ingredient consists of synthetic (crystalline) lycopene to be marketed in three different formulations. These are lycopene 10 %, lycopene 10 cold water dispersion (CWD) and lycopene dispersion 20 %.

Synthetic lycopene is suggested by the applicant to be used in food supplements at levels of 8 or 15 mg/dosing, in beverages and dairy products at levels of up to 2.5 mg/100 g, in breakfast cereals up to 4 mg/100 g, in cereal bars up to 8 mg/100 g, in fats and dressings up to 4 mg/100 g and in dietary foods for special medical purposes at levels in accordance with the particular requirements of the person for whom the products are intended.

The applicant provides an intake estimate of lycopene based on three sources including 1) normal dietary intake from food, 2) intake from dietary supplements and 3) intake from proposed fortified food products. The Panel notes that an additional source is 4) use as a food colour.

[1] For citation purposes: Scientific Opinion of the Panel on Panel on Dietetic Products, Nutrition and Allergies on a request from the European Commission on the safety of synthetic lycopene. *The EFSA Journal* (2008) 676, 1-25



Safety of Synthetic Lycopene

TABLE OF CONTENTS

Group E

SCIENTIFIC OPINION

Safety of isomaltulose syrup (dried) as a novel food pursuant to Regulation (EU) 2015/2283

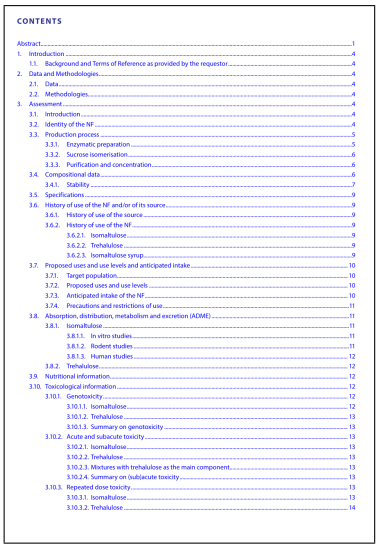
EFSA Panel on Nutrition, Novel Foods and Food Allergens (NDA) | Dominique Turck | Torsten Bohn | Jacqueline Castenmiller | Stefaan De Henauw | Karen Ildico Hirsch-Ernst | Alexandre Maciuk | Inge Mangelsdorf | Harry J. McArdle | Androniki Naska | Kristina Pentieva | Alfonso Siani | Frank Thies | Sophia Tsabouri | Marco Vinceti | Margarita Aguilera-Gómez | Francesco Cubadda | Thomas Frenzel | Marina Heinonen | Rosangela Marchelli | Monika Neuhäuser-Berthold | Morten Poulsen | Miguel Prieto Maradona | Josef Rudolf Schlatter | Alexandros Siskos | Henk van Loveren | Lucien Ferreira da Costa | Océane Albert | Helle Katrine Knutsen

Abstract

Following a request from the European Commission, the EFSA Panel on Nutrition, Novel Foods and Food Allergens (NDA) was asked to deliver an opinion on isomaltulose syrup (dried) as a novel food (NF) pursuant to Regulation (EU) 2015/2283. The NF consists of a mixture of mono- and disaccharides in powder form, mainly composed of isomaltulose (≥ 75%) and trehalulose (< 13%). The applicant intends to use the NF as a replacement for sucrose already on the market. The information provided on the manufacturing process, composition and specifications of the NF is sufficient and does not raise safety concerns. No absorption, distribution, metabolism and excretion (ADME) or toxicological data were provided for the NF. Instead, the safety of the NF was assessed based on literature data available on isomaltulose and mixtures of isomaltulose and trehalulose. In addition, considering the nature, compositional characterisation and production process of the NF, the Panel considered that such data were sufficient to conclude that the NF is as safe as sucrose.

KEYWORDS

isomaltulose; isomaltulose syrup; novel foods; trehalulose



CONTENTS

The distribution in the groups is as follows:

group	count
A	142
B	32
C	4
D	14
E	6

4 Extraction rules

- rozdelit na casti, ktere lze ziskat z FRONT

4.1 DOI

The DOI can be found in article-meta - article-id - pub-id-type="doi".

```
<front>
    <article-meta>
        <article-id pub-id-type="doi">10.2903/j.efsa.2023.7904</article-id>
```

4.2 Adoption date

The traditional foods do not have adoption date, instead they have approval date, which is not included in the JATS files and therefore cannot be extracted. The novel foods have adoption date.

front - notes -fn-group - v nejake z nich je Adopted group6 - ADOPTED: in fn-group

```
<front>
    <notes>
        <fn-group>
            <fn id="efs26305-note-1203" xml:lang="en">
                <p>Adopted: 22 October 2020</p>
            </fn>
        </fn-group>
    </notes>
</front>
```

4.3 Publication Date

Publication date can be found in article-meta - pub-date - pub-type="epub".

4.4 Question Number

question number je taky v jedne fn-group

```
<front>
    <notes>
        <fn-group>
            <fn id="efs26305-note-1002" xml:lang="en">
                <p>
                    <bold>Question number:</bold>
                    EFSA-Q-2020-00491
                </p>
            </fn>
        </fn-group>
    </notes>
</front>
```

4.5 Scientific Panels

Usually NDA, for traditional foods EFSA, sometimes also GMO.

group 3 - v title: Panel on dietetic products, nutrition and allergies group 6 - article-meta : contrib-group : collab(collab-type="authors") : NDA

look into collab - some panel there? (gmo, nda) -¿ added - no panel there? - look into title -¿some panel there -¿ added -¿ no panel there -¿ panel='EFSA'

ve formátu **Group 2** je to až někde úplně dole: v body → pak najít sekci s tímto titlem

```
<sec id="efs28416-sec-0024" xml:lang="en">
    <title>QUESTION NUMBER</title>
    <p xml:lang="en">EFSA-Q-2020-00491</p>
</sec>
```

pro nalezeni pouzity regex:

```
r 'EFSA[-]Q[-]\S+'
```

4.6 Scientific Officer

EFSA provided us with list of possible scientific officers. vic popsat jak toto funguje

4.7 Mandate type

–order matters

- guidance - always has guidance in the title - traditional food - always have notification of XX as traditional food
- extension of use - new dossier - the remaining ones
- nutrient source - can be added

4.8 Novel Food Name

- v title v article meta → article-title

4.9 Category

One NF can fall under multiple categories. Categories as per Regulation Article 3 of 2015/2283. Traditional food guidelines specify following categories

Categories are defined in specific european regulations. Each regulation defines its own set of categories.

REGULATION (EC) No 258 /97 from 27 January 1991

(a) foods and food ingredients containing or consisting of genetically modified organisms within the meaning of Directive 90 /220 /EEC;

(b) foods and food ingredients producted from, but not containing, genetically modified organisms;

(c) foods and food ingredients with a new or intentionally modified primary molecular structure;

(d) foods and food ingredients consisting of or isolated from micro-organisms, fungi or algae;

(e) foods and food ingredients consisting of or isolated from plants and food ingredients isolated from animals, except for foods and food ingredients obtained by traditional propagating or breeding practices and having a history of safe food use;

(f) foods and food ingredients to which has been applied a production process not currently used, where that process gives rise to significant changes in the composition or structure of the foods or food ingredients which affect their nutritional value, metabolism or level of undesirable substances .

commission recommendation 97/618/EC from 29 July 1997

(a) pure chemicals or simple mixtures from non-GM sources;

(b) complex NF from non-GM source;

(c) GM plants and their products;

(d) GM animals and their products;

(e) GM microorganisms and their products;

(f) foods produced using a novel process.

REGULATION (EU) 2015/2283 from 25 November 2015, Article 3

(i) food with a new or intentionally modified molecular structure, where that structure was not used as, or in, a food within the Union before 15 May 1997;

(ii) food consisting of, isolated from or produced from microorganisms, fungi or algae;

(iii) food consisting of, isolated from or produced from material of mineral origin;

(iv) food consisting of, isolated from or produced from plants or their parts, except when the food has a history of safe food use within the Union and is consisting of, isolated from or produced from a plant or a variety of the same species obtained by:

— traditional propagating practices which have been used for food production within the Union before 15 May 1997; or

— non-traditional propagating practices which have not been used for food production within the Union before 15 May 1997, where those practices do not give rise to significant changes in the composition or structure of the food affecting its nutritional value, metabolism or level of undesirable substances;

(v) food consisting of, isolated from or produced from animals or their parts, except for animals obtained by traditional breeding practices which have been used for food production within the Union before 15 May 1997 and the food from those animals has a history of safe food use within the Union;

(vi) food consisting of, isolated from or produced from cell culture or tissue culture derived from animals, plants, micro-organisms, fungi or algae;

(vii) food resulting from a production process not used for food production within the Union before 15 May 1997, which gives rise to significant changes in the composition or structure of a food, affecting its nutritional value, metabolism or level of undesirable substances;

(viii) food consisting of engineered nanomaterials as defined in point (f) of this paragraph;

(ix) vitamins, minerals and other substances used in accordance with Directive 2002/46/EC, Regulation (EC) No 1925/2006 or Regulation (EU) No 609/2013, where:

— a production process not used for food production within the Union before 15 May 1997 has been applied as referred to in point (a) (vii) of this paragraph; or

— they contain or consist of engineered nanomaterials as defined in point (f) of this paragraph;

(x) food used exclusively in food supplements within the Union before 15 May 1997, where it is intended to be used in foods other than food supplements as defined in point (a) of Article 2 of Directive 2002/46/EC;

5 Composition extraction

why was deemed important how was performed - gpt assisstent + query success rate

6 Implementation

-mozna neni potreba zas tak popisovat Two main classes:

JATS Opinion

- encompass the formal parts of the opinion, abstract the inner structure of the opinions into common interface

- it will abstract the formal things about the article: DOI, Title, EFSA Question, adoption date, publication date, URL, authors(panels)

- also it will abstract the different headings from different types of opinion and match them to the same structure

Opinion

- the scientific opinion itself, it will accept JATS opinion in constructor

- it will have subclass for each category, which will contain the category specific information

- it will extract NF Name, Applicant, Country of origin, NF code, type of mandate, regulation, outcome, common name, trade names, food form
- also proposed uses + target population
- availability of ADME studies
- allergenicity
- food category – for FOODS
- for **Microorganisms**: type, genus, species, strain, QPS(?)
- for **Plants**: type, common name, botanical name, genus, species, part used
- for **Animals**: genus, species, subspecies, part used
- for **Cell or Tissue**: genus, species, cell type
- for **Chemicals**: common name, IUPAC name, CAS notation, SMILES, molecular formula, InChi