European Commission

JRC TECHNICAL REPORTS

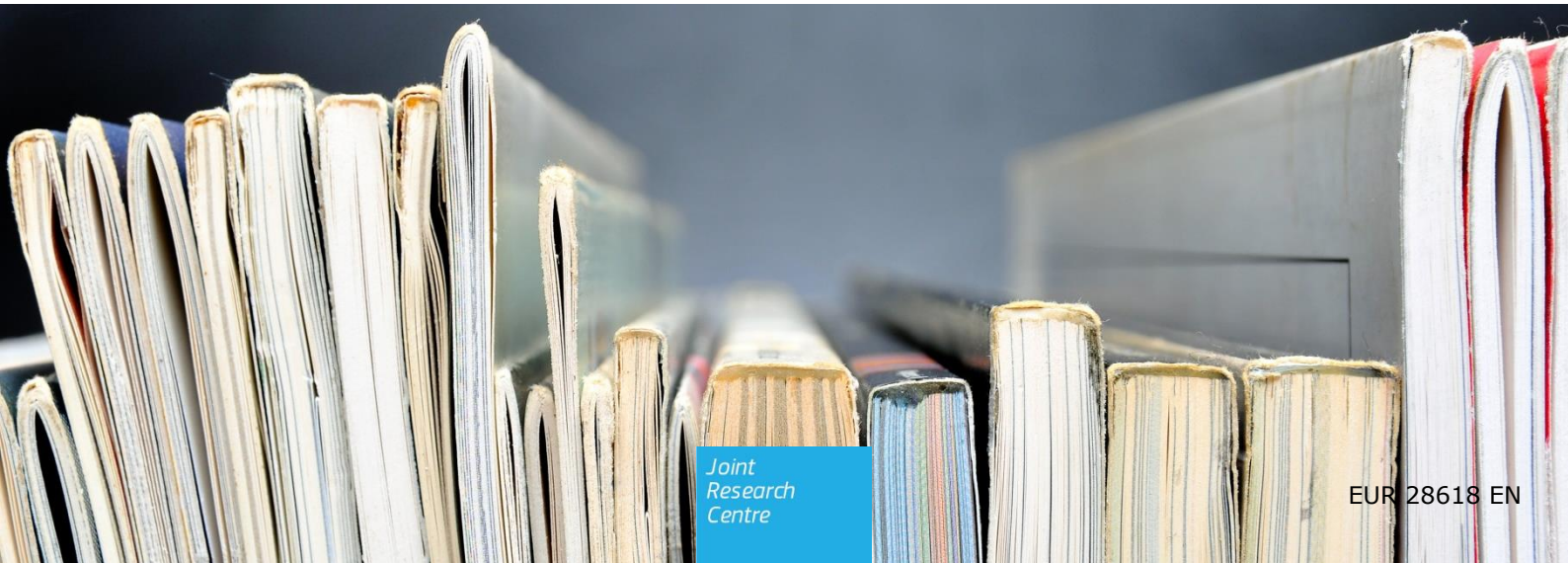
Clustering and classification of reference documents from large-scale literature searches

Support to the SAM explanatory note "New Techniques in Agricultural Biotechnology"

Angers, Alexandre

Petrillo, Mauro

2017



Joint Research Centre

EUR 28618 EN

JRC Science Hub
https://ec.europa.eu/jrc

JRC106191

EUR 28618 EN

PDF ISBN 978-92-79-69019-8 ISSN 1831-9424 doi: 10.2760/256470

Luxembourg: Publications Office of the European Union, 2017

How to cite this report: Angers A. and Petrillo M., *Clustering and classification of reference documents from large-scale literature searches*, EUR 28618, doi: 10.2760/256470

Contents

1 Acknowledgements

We would like to warmly thank our colleagues from DG RTD Dulce Boavida, Sigrid Weiland, Jeremy Bray and Stuart Kirk for their invaluable input to this report.

2 Introduction

Searches of the existing scientific literature are the cornerstone of scientific research and reporting and often play a key role in the provision of scientific advice to inform policy and practice. With the constant growth of the rate at which scientific studies and reviews are being published, the information produced by these searches can be challenging to manage. Narrowing the search increases the risk of overlooking important documents, and/or introducing bias, while broadening it can produce too many documents to be reasonably processed.

This report describes a set of strategies designed to process large sets of scientific references (such as those obtained by broad literature searches, in different databases) and assist in the identification of documents relevant for specific aspects.. These strategies take advantage of metadata associated to each document in SCOPUS, the database of peer-reviewed literature maintained by Elsevier and accessible through an Application Programming Interface (API).

These strategies were developed in collaboration with the colleagues from RTD SAM UNIT and applied in support to the European Commission's Scientific Advice Mechanism's High Level Group (SAM HLG) of Scientific Advisors in managing and clustering and assigning the results of literature searches in the context of the development of the group's explanatory note "*New Techniques in Agricultural Biotechnology*"[1]. The explanatory note, which was published in April 2017, required the compilation and screening of a large amount of scientific publications within a short time scale, from which the note was subsequently developed. In the final selection of references, greater emphasis was placed upon review articles, opinions and reports.

[1] https://ec.europa.eu/research/sam/index.cfm?pg=agribiotechnology

3 The SCOPUS database

The main resource used to process the set of reference documents was SCOPUS database, developed and maintained by Elsevier (Burnham, 2006). This choice was motivated by a few important characteristics of this database, including:

1. The capacity to automatically query the database through scripts. The SCOPUS Application program interface (API) is publicly available and well documented[2]
2. The possibility to receive the results of queries in a "machine-friendly" format such as XML (Bray et al., 1998) and JSON (Crockford, 2006). For our purposes, the JSON format was used.
3. The extensive metadata associated to each record in the SCOPUS database. This includes the full title, the abstract, the list of keywords and the complete list of articles cited in the text of the record.

One of the reported weaknesses of the SCOPUS database, i.e. the fact that it is limited to recent articles (published after 1995) (Falagas et al., 2007), is not a concern as a decision was made by the SAM HLG to limit the searches to references published after 2003.

For the New Techniques in Agricultural Biotechnology explanatory note the SAM secretariat chose Mendeley as a reference management and collaborative tool to compile, organise and annotate the project's research citations (Holt Zaugg, Richard E. West, Isaku Tateishi, Daniel L. Randall, 2011).

The first step of reference processing was then the matching process of all the compiled records, from SCOPUS, BIOSIS, Find-eR and Web of Science (WoS) to their equivalent within the SCOPUS database.

The details of how this was achieved are described in Annex I.

This step caused the loss of some of the compiled references, in case the record did not exist in SCOPUS or could not be successfully linked. This loss remained limited, as 2,301 of the 2,727 references (84%) could be successfully identified and used for further analyses. This process helped also on the identification of duplicates and later on references that were out of scope.

[2] https://dev.elsevier.com/

4 Bibliography clustering

The first strategy to extract useful information from a large set of reference articles involved a comparison of the references cited by each of these articles in their bibliography. In particular, the goal was to highlight "clusters" of articles that overlap significantly in the other literature they cite.

This concept is illustrated in Figure 1:

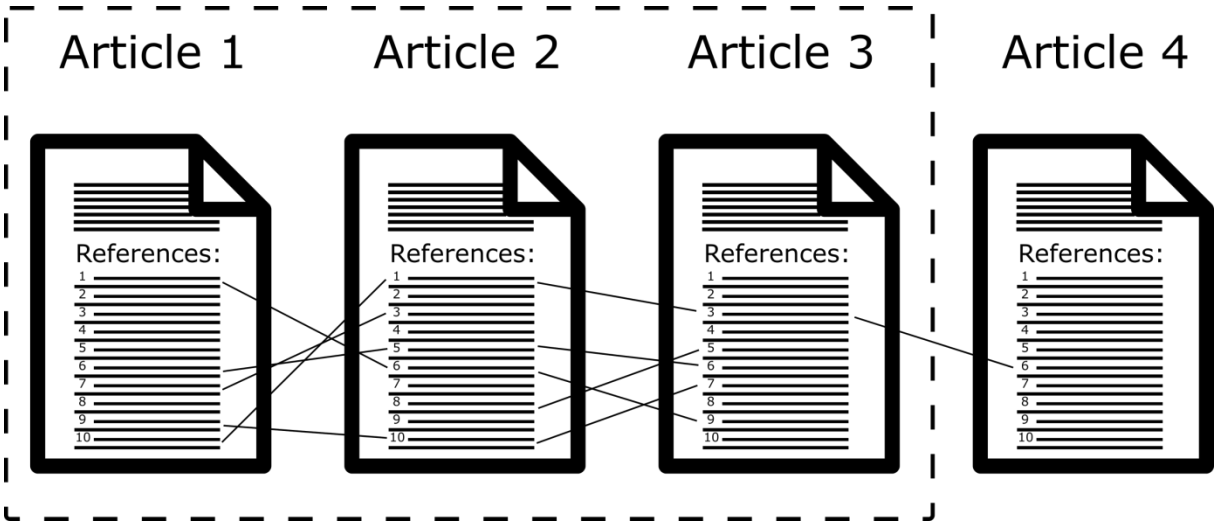


Figure 1: Concept of the biobliography clustering

In this illustration, Articles 1, 2 and 3 are part of a cluster, as a significant number of the articles in their bibliographies are identical. Article 4, on the other hand, only cites 1 article present in the reference list of Articles 1-3, so is not part of this cluster.

Once applied to a large set of references, this clustering can be used for different purposes. These include:

1. The identification of "redundant" articles. This is particularly true when the threshold of common references is set high (e.g. 50% and more). If the overlap between the documents cited by two articles is very high, it might not be necessary to read both, in particular when the number of other articles to consider is high.
2. Conversely, at lower thresholds, the clustering can be used to find supplementary references to support a specific statement or identify differing views of the same subject. Using a known reference for a specific section of the document being drafted, it can useful to verify if this reference is part of a "cluster", as the other articles in this cluster would most probably cover the same specific subject, if they cite the same literature.

The technical details on how this was done, in practice, are found in Annex II.

This clustering exercise was performed on the 2,301 references for the "New Techniques in Agricultural Biotechnology" that were found in SCOPUS, with a threshold of 25% common references.

With this set of articles and threshold, 129 clusters were formed. Their size distribution is shown in Figure 2.

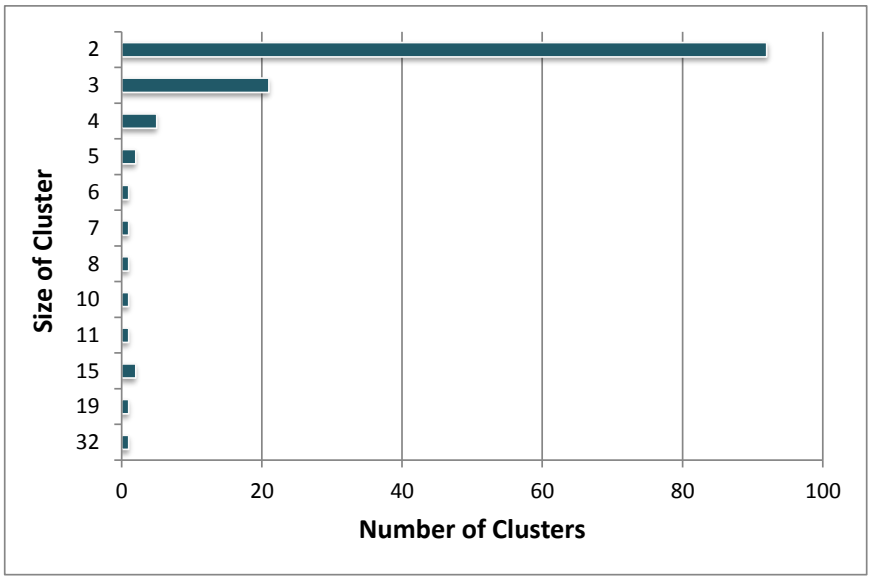


Figure 2: Size distribution of the clusters formed (vertical axis), showing the number of clusters formed of this size (horizontal axis)

This distribution shows that the majority of clusters are formed of 2 or three articles, although single "big" clusters (like one composed of 32 articles) were identified.

Analysis of some of the clusters allowed some interesting observations. For example, the following cluster is composed of 5 articles, and shows a clear theme of gene editing in animals, and in particular pigs:

Authors	Title	Journal	Date of publication
Tan W.S. et al	Precision Editing of Large Animal Genomes	Advances in Genetics	25/10/2012
Whyte J.J. et al	Genetic modifications of pigs for medicine and agriculture	Molecular Reproduction and Development	01/10/2011
Laible G. et al	Improving livestock for agriculture - technological progress from random transgenesis to precision genome editing heralds a new era	Biotechnology Journal	01/01/2015
Zhang X. et al	Advances in the Generation of Transgenic Domestic Species via Somatic Cell Nuclear Transfer	Principles of Cloning: Second Edition	01/10/2013
Whyte J.J. et al	Cell biology symposium: Zinc finger nucleases to create custom-designed modifications in the swine (Sus scrofa) genome	Journal of Animal Science	01/04/2012

It is interesting to note that some of the clusters are composed of articles sharing authors, and in particular first authors, as for the following cluster:

Authors	Title	Journal	Date of publication
Niemann H. et al	Somatic Cloning and Epigenetic Reprogramming in Mammals	Principles of Regenerative Medicine	01/12/2011
Niemann H. et al	Epigenetic reprogramming in mammalian species after SCNT-based cloning	Theriogenology	01/07/2016
Niemann H. et al	Epigenetic reprogramming in embryonic and foetal development upon somatic cell nuclear transfer cloning	Reproduction	01/02/2008

The presence of these clusters can be used an internal control for the clustering script, as these will most probably be written by the same scientist(s) and thus are the most likely to share a common bibliography.

5 Classification by keywords (assignment)

5.1 Process

In the early stages of planning with regards to the structure and content of the SAM HLG document, "search grids" were prepared. These grids listed key concepts in order to identify the supporting literature retrieved from the search. The content of these search grids generally mirrored structure different sections of the of the explanatory note.

In order to map the results of the literature searches to the different cells of these search grids (assignments), keywords were identified that were in turn fed into *in-house* developed scripts to map, when possible, each reference to specific cells of the search grids corresponding to different sections of the explanatory note.

The different concepts present in the search grids were separated into three different "dimensions" for the final categorisation of articles: **Technologies** (e.g. oligonucleotide directed mutagenesis or transgenesis), **Applications** (e.g. plants or animals) and **Aspects** (e.g. cost or safety).

For each of the keywords, specific search terms were identified to be used for the assignment.

The selected search terms for each keyword were:

Technologies	
Keywords	Search term(s)
Conventional breeding techniques	
Sexual crosses	'sexual cross'
Bridge crosses	'bridge cross'
Embryo rescue	'embryo rescue'
Somatic hybridisation	'somatic hybridisation' or 'somatic hybridization'
Translocation breeding	'translocation breeding'
Mutation breeding	'mutation breeding'
Simple selection	'simple selection'
Somaclonal variation	'somaclonal variation'
Cell selection	'cell selection'
Hybridization	'hybridization' or 'hybridisation'
Polyploidy induction	'polyploidy induction'
Natural mating	'natural mating'
Artificial insemination from selected sires	'artificial insemination'

Oocyte collection from selected dams	'ocyte collection'[3]
Embryo selection and transfer from selected genitors	'embryo selection'
In vitro fertilisation	'in vitro fertilisation' or 'in vitro fertilization'
Long term storage of gametes and embryos	'long term storage' or 'cryopreservation'
Embryo splitting	'embryo splitting'
Mutagenesis and selection	'mutagenesis'
Conjugation and selection	'conjugation'
Natural transduction	'natural transduction
Natural transformation	'natural transformation'
Protoplast fusion	'protoplast fusion'
Established techniques of genetic modification in biotechnology	
Insertion of nucleic acid molecules into a host organism	'transgen'
Introduction of heritable material, including micro-injection, macro-injection and micro-encapsulation	'injection' or 'incapsulation'
Cell fusion or hybridisation techniques	'cell fusion'
New techniques of genetic modification in biotechnology	
Oligonucleotide directed mutagenesis	'oligonucleotide directed mutagenesis' or 'odm'
Zinc finger nuclease	'zinc finger' or 'zinc-finger' or 'zfn'
Cisgenesis	'cisgenesis'
Intragenesis	'intragenesis'
Agro-infiltration	'agro-infiltration' or 'agroinfiltration'
RNA-dependent DNA methylation	'rna-dependent dna methylation' or 'rddm'
Reverse breeding	'reverse breeding'
Recent genome editing technologies	'crisp' or 'cas9'
	'talen'

[3] to cover for both oocyte, ocyte and ovocyte

Applications	
Keyword	Search term(s)
Plants	'plant' or 'crop' or 'forage'
Animals	'animal' or 'livestock' or 'fish' or 'insect'
Micro-organisms	'microbe' or 'micro-organism' or 'yeast' or 'saccharomyces'
Gene drive	'gene drive'
Synthetic biology	'synthetic biology'

Aspects	
Keyword	Search term(s)
Molecular Mechanism	'mechanism' or 'specificit'
Safety for health	'health' or 'biosafety'
Safety for environment	'environment' or 'biosafety'
Detectability of products	'detect'
Speed	'speed'
Costs	'cost' or 'price'
Maturity	'maturity'

These terms were then searched, as a single text matching, in the title, abstract and MeSH[4] keywords (if available) of the articles, ignoring case. Any hit assigned the article to the corresponding category. The final categories were composed of the three dimensions (3D) described above, i.e.

An **Aspect** (3rdD) relating to a **Technology** (1stD) used in a specific **Application** (2ndD).

Each article could be assigned to more than one category, as long as the search terms were present. For each dimension, an "unassigned" option was added, for example when an article could be assigned as covering the safety (aspect) of transgenesis (technology), with no mention of a specific application.

The technical details on how this was done, in practice, are found in Annex III.

[4] MeSH (Medical Subject Headings) is the National Library of Medicine controlled vocabulary thesaurus used for indexing articles for PubMed

After categorisation of the articles in Technologies, Applications and Aspects, the results were formatted in Excel. The following 'hierarchical' scheme was used:

<table>
<tr><td rowspan="6">Technology 1</td><td rowspan="4">Application 1</td><td rowspan="3">Aspect 1</td><td>Article 1</td></tr>
<tr><td>Article 2</td></tr>
<tr><td>…</td></tr>
<tr><td>Aspect 2</td><td></td></tr>
<tr><td>…</td><td></td></tr>
<tr><td>Application 2</td><td></td><td></td></tr>
<tr><td></td><td>…</td><td></td><td></td></tr>
<tr><td>Technology2</td><td></td><td></td><td></td></tr>
<tr><td>…</td><td></td><td></td><td></td></tr>
</table>

In addition, XML files were produced that allowed the generation of nested folders in Mendeley through the "import" feature, each containing the classified references.

A sample of the final structure is shown in Figure 3.

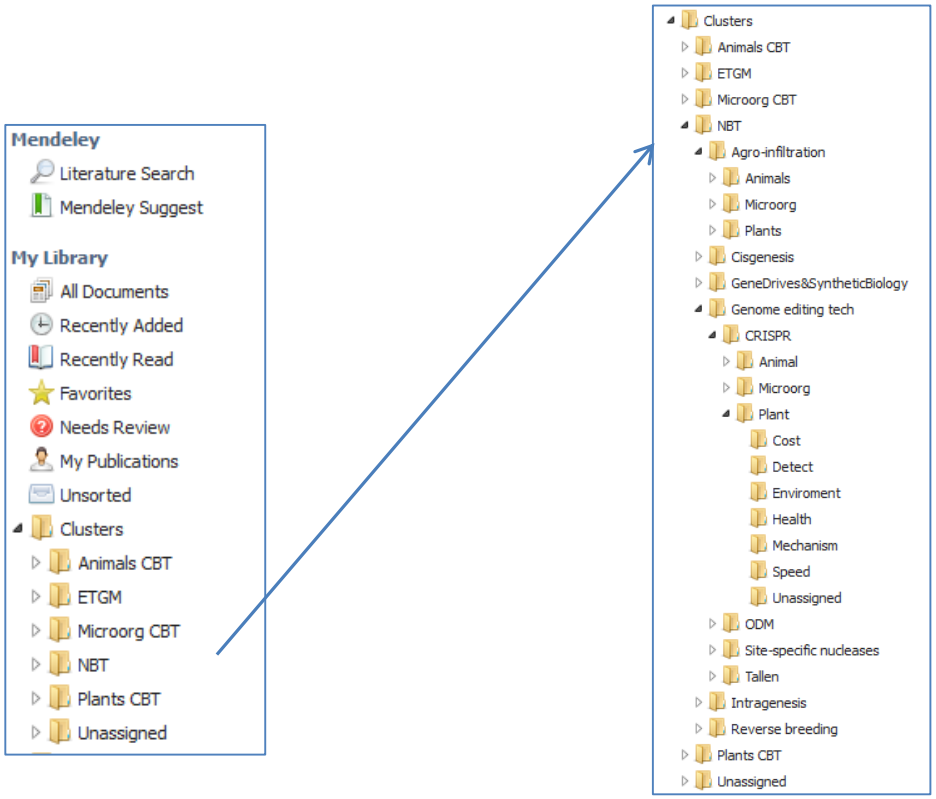


Figure 3: Screenshot – Mendeley clusters

5.2 Comparison with categorisation by experts

Prior to the launch of the scientific literature searches for the explanatory note, topic experts were asked to draft a list of key references. In addition, and of interest for this work, they were asked to assign each of these key references to one or more sections planned for the final document.

For these key references, the results of the automatic assignment by our scripts were compared to this expert assignment, with the results shown in the following table:

Authors	Title	Script assignment	Experts assignment
Altpeter F. et al	Advancing crop transformation in the era of genome editing	No technique / Micro-organisms, No technique / Synthetic biology	Insertion of nucleic acid molecules into a host organism – molecular mechanism
Andersen M.M. et al	Feasibility of new breeding techniques for organic farming	No technique / Plants	-
Barrangou R. et al	Applications of CRISPR technologies in research and beyond	CRISPR/Cas9 / Plants, CRISPR/Cas9 / Animals, CRISPR/Cas9 / Gene drive	Recent genome editing technologies, Comparisons new techniques and conventional breeding techniques, Comparisons new techniques and established techniques of genetic modification in biotechnology
Barrangou R. et al	Exploiting CRISPR-Cas immune systems for genome editing in bacteria	CRISPR/Cas9 / Micro-organisms / Mechanism	Micro-organisms (safety, speed, costs, maturity, detectability)
Bortesi L. et al	Patterns of CRISPR/Cas9 activity in plants, animals and microbes	CRISPR/Cas9 / Plants, CRISPR/Cas9 / Animals, CRISPR/Cas9 / Micro-organisms	Recent genome editing technologies
Elling U. et al	Genome wide functional genetics in haploid cells	CRISPR/Cas9 / Animals, CRISPR/Cas9 / Micro-organisms	-
Hauschild-Quintern J. et al	Gene knockout and knockin by zinc-finger nucleases: Current status and perspectives	Transgenics / Plants / Mechanism, Transgenics / Animals / Mechanism, Zinc finger nuclease / Plants / Mechanism, Zinc finger nuclease / Animals / Mechanism	Recent genome editing technologies, Zinc Finger Nuclease
Kues W.A. et al	Reproductive biotechnology in farm animals goes genomics	Artificial Insemination / Animals / Cost, Transgenics / Animals / Cost, Micro-injection, macro-injection and micro-encapsulation / Animals / Cost	Conventional breeding techniques in animals
Kues W.A. et al	Advances in farm animal transgenesis	Zinc finger nuclease / Animals / Safety for environment	New techniques
Liu D. et al	Advances and perspectives on the use of CRISPR/Cas9 systems in plant genomics research	CRISPR/Cas9 / Plants, CRISPR/Cas9 /Synthetic biology	Recent genome editing technologies

Liu W. et al	Plant synthetic promoters and transcription factors	Transgenics / Plants / Mechanism	Insertion of nucleic acid molecules into a host organism, Recent genome editing technologies, Synthetic genomics
Liu W. et al	Advanced genetic tools for plant biotechnology	No technique / Plants / Environmental safety	Insertion of nucleic acid molecules into a host organism, Recent genome editing technologies
Liu W. et al	Plant synthetic biology	No technique / Plants , No technique / Synthetic biology	Insertion of nucleic acid molecules into a host organism, Recent genome editing technologies, Synthetic genomics
Luo M. et al	Applications of CRISPR/Cas9 technology for targeted mutagenesis, gene replacement and stacking of genes in higher plants	Mutagenesis / Plants / Mechanism, Zinc finger nuclease / Plants / Mechanism, CRISPR/Cas9 / Plants / Mechanism, TALENs / Plants / Mechanism	Recent genome editing technologies – molecular mechanism / products
Lusser M. et al	Comparative regulatory approaches for groups of new plant breeding techniques	No technique / Plants	-
Ma X. et al	CRISPR/Cas9 Platforms for Genome Editing in Plants: Developments and Applications	Zinc finger nuclease / Plants / Mechanism, CRISPR/Cas9 / Plants / Mechanism, TALENs / Plants / Mechanism	Recent genome editing technologies
Mougiakos I. et al	Next Generation Prokaryotic Engineering: The CRISPR-Cas Toolkit	CRISPR/Cas9 / Micro-organisms / Mechanism	-
Niemann H. et al	Perspectives for feed-efficient animal production	Transgenics / Animals / Safety for environment	New techniques and conventional breeding techniques, New techniques and established techniques of genetic modification in biotechnology
Niemann H. et al	Epigenetic reprogramming in mammalian species after SCNT-based cloning	No technique / Animals	Conventional breeding techniques in animals
Niemann H. et al	Somatic Cloning and Epigenetic Reprogramming in Mammals	Unassigned	Conventional breeding techniques in animals
Niemann H. et al	Epigenetic reprogramming in embryonic and foetal development upon somatic cell nuclear transfer	Unassigned	Conventional breeding techniques in animals

	cloning		
Petersen B. et al	Molecular scissors and their application in genetically modified farm animals	Transgenics / Plants / Mechanism, Transgenics / Animals / Mechanism, Zinc finger nuclease / Plants / Mechanism, Zinc finger nuclease / Animals / Mechanism, CRISPR/Cas9 / Plants / Mechanism, CRISPR/Cas9 / Animals / Mechanism, TALENs / Plants / Mechanism, TALENs / Animals / Mechanism	Zinc Finger Nuclease
Schaart J.G. et al	Opportunities for Products of New Plant Breeding Techniques	No technique / Plants	Recent genome editing technologies
Schiml S. et al	Revolutionizing plant biology: Multiple ways of genome engineering by CRISPR/Cas	CRISPR/Cas9 / Plants	Recent genome editing technologies
Urrego R. et al	Epigenetic disorders and altered gene expression after use of assisted reproductive technologies in domestic cattle	Cryopreservation / Animals	Conventional breeding techniques in animals
Wright A.V. et al	Biology and Applications of CRISPR Systems: Harnessing Nature's Toolbox for Genome Engineering	CRISPR/Cas9 / Plants / Mechanism, CRISPR/Cas9 / Animals / Mechanism, CRISPR/Cas9 / Micro-organisms / Mechanism	Recent genome editing technologies, Others - special view to applications for gene drives and in synthetic biology
Zhu C. et al	Characteristics of Genome Editing Mutations in Cereal Crops	Zinc finger nuclease / Plants , CRISPR/Cas9 / Plants	Recent genome editing technologies – molecular mechanism/ products

This table shows a general consistency between the automatic assignment of categories and the expert assignment to specific EN sections. At times, the automatic assignment was more exhaustive, such as for the article by Petersen et al., "Molecular scissors and their application in genetically modified farm animals", assigned by experts to the section on Zinc Finger Nucleases, but by the script to all the different "molecular scissors", including ZFNs but also CRISPR/Cas9 and TALENs. On the other hand, a set of three references by Niemann et al. on somatic cell nuclear transfer cloning were mostly unassigned (at best, correctly linked to an "animals" application), since this technique was not part of the keywords selected for the conventional breeding techniques.

5.3 Visualisation of results

In order to understand the coverage of the different categories, the number of articles assigned to each Technology was counted. This number was then subdivided by Application and Aspect. The results are shown in Figure 4 below.

The results show that the techniques of hybridisation, mutagenesis, transgenesis and Crisp/Cas9 are well covered for all applications and all aspects. On the other hand, for "embryo rescue", only few articles were found. Techniques to which no articles were assigned are not included in the graph.

This figure can be helpful to evaluate potential gaps in the search results obtained.

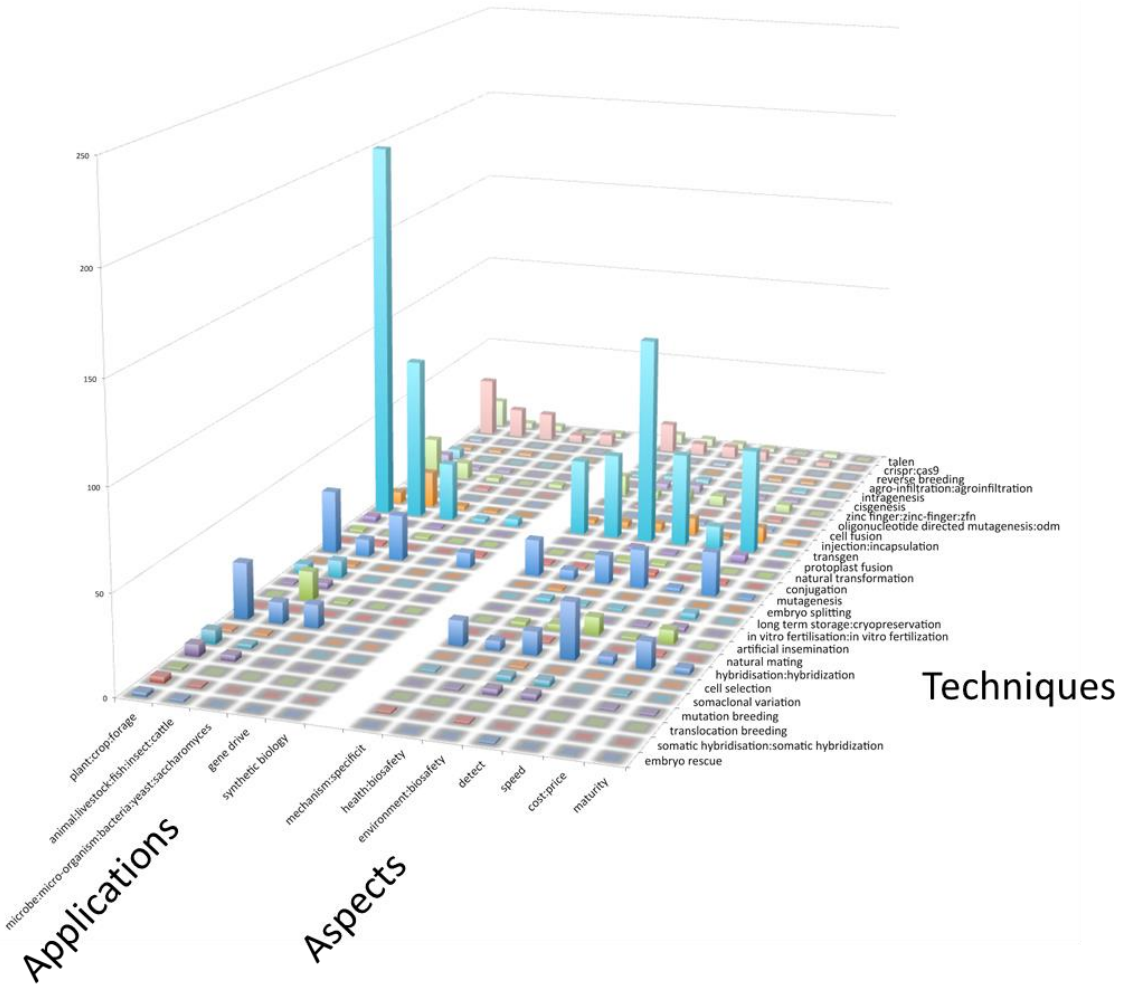


Figure 4: Distribution of the automatic assignment of articles to the different keywords

Another visualisation is to evaluate the profile of individual articles, in terms of how many techniques, applications and aspects they cover (based on our categorisation). Figure 5 provides an idea of the percentage of applications (total possible: 5), techniques (total possible: 34) and aspects (total possible: 7) "dimensions" covered by the individual articles. The size of the bubbles (as fourth dimension) indicates how many articles cover a specific combination of the three dimensions.

As expected, individual articles cover only a minority of techniques (maximum: 7/34, about 20%). It is more common for articles to cover many of the different aspects or applications of these techniques. The red sphere represents the average for the four dimensions.

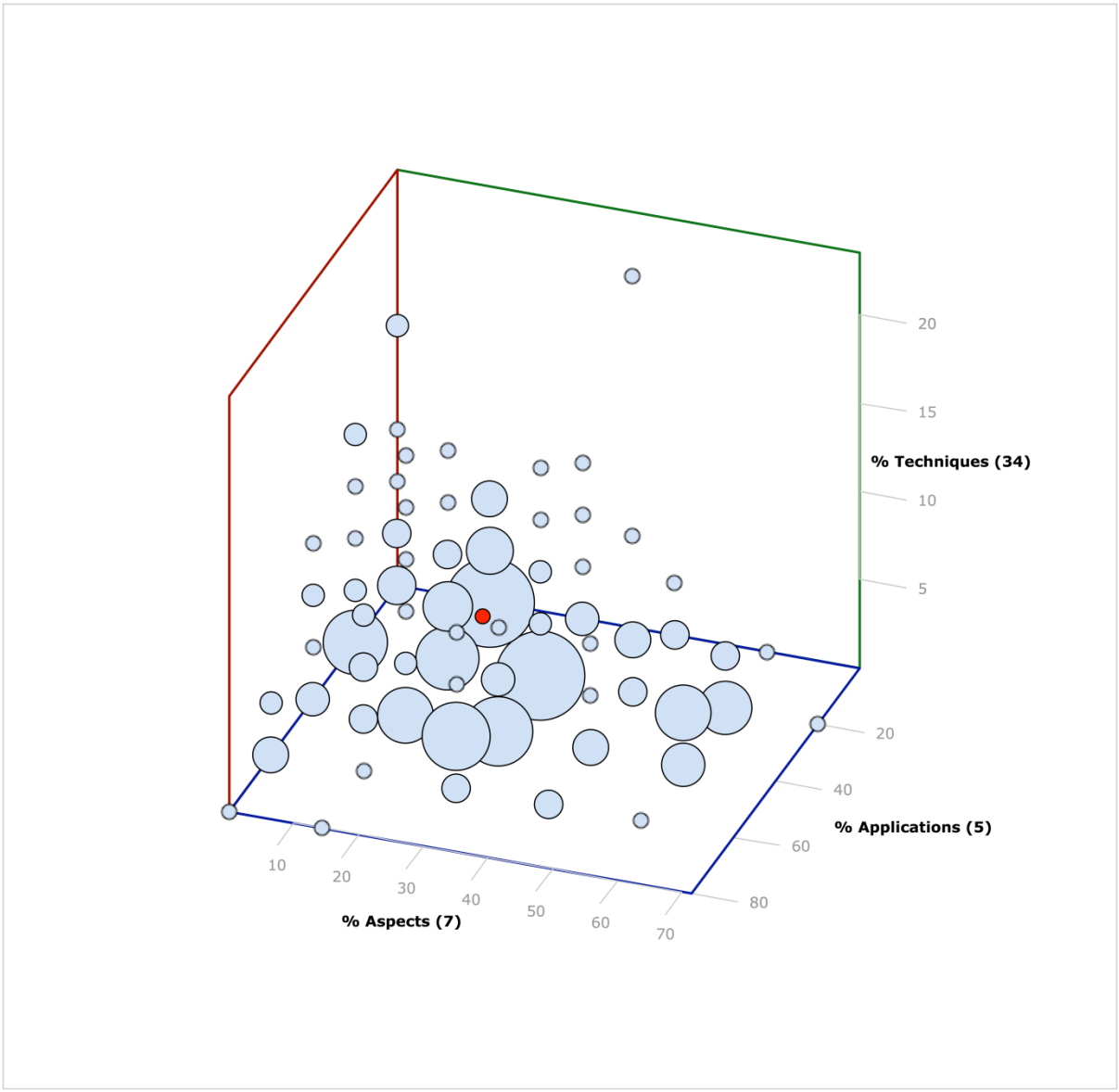


Figure 5: Distribution of individual articles, in terms of how many techniques, applications and aspects they were assigned to.

6 Conclusions

Finding relevant scientific information in the published literature is an ever-growing challenge, as the amount of articles and reviews published increase constantly. Broad searches can produce a large amount of hits that can be difficult to process. On the other hand, focused searches can produce manageable numbers of results, with the risk of missing key references and introducing bias depending on the terminology used.

The strategies presented within this report were found to be effective and efficient in identifing *a posteriori* important references from a large pool of references from a literature search. Clustering strategies (section 3) were very useful in highlighting groups of references relevant to the same concept, and were used to expand on a "core" of known articles. Assignment strategies (section 4) allowed the rapid identification of specific references relevant to different sections of the draft explanatory note, based on an initial map of the concepts to be investigated.

Overall these strategies proved to be a highly practical and valuable aid in the development of science advice documents based on the screening and processing of large quantities of scientific literature, in particular by facilitating the grouping of the identified literature according to aspects and relevance, the identification of gaps in the available evidence and the identification of further, complementary evidence sources.

New tools and databases are constantly being developed and made available, processing the existing and newly published scientific literature in order to identify and highlight important information, trends, metadata, etc. It is important to use this new information to improve the research process involved in producing new scientific, review, and state-of-the-art articles, using strategies such as those described in the current report.

References

Bray, T., Paoli, J., Sperberg-McQueen, C.M., Maler, E., and Yergeau, F. (1998). Extensible markup language (XML). World Wide Web Consort. Recomm. REC-Xml-19980210 Httpwww W3 OrgTR1998REC-Xml-19980210 *16*, 16.

Burnham, J.F. (2006). SCOPUS database: a review. Biomed. Digit. Libr. *3*, 1.

Crockford, D. (2006). The application/json media type for javascript object notation (json).

Falagas, M.E., Pitsouni, E.I., Malietzis, G.A., and Pappas, G. (2007). Comparison of PubMed, SCOPUS, Web of Science, and Google Scholar: strengths and weaknesses. FASEB J. *22*, 338–342.

Holt Zaugg, Richard E. West, Isaku Tateishi, Daniel L. Randall (2011). Mendeley: Creating Communities of Scholarly Inquiry Through Research Collaboration. TechTrends *55*, 32–36.

List of figures

Annexes

Annex 1. Identification of the relevant references in SCOPUS

The results of a scientific literature search performed by a Review Team on four separate platforms/databases were compiled in a folder on the Mendeley platform. The role of the Review Team was primarily to find and collate the information/evidence upon which the Explanatory Note was produced. The Review Team operated under the direction of a Steering Group which conducted the bulk of the evidence syntheses to produce the Explanatory Note.

In order to perform the analyses described in the current report, the equivalent records in SCOPUS were identified using the following steps (executed as part of *in-house* developed scripts in the Ruby language):

1. A list of references in the XML format was generated using the "Export" function of Mendeley (Endnote format).
2. Within this file, the DOIs of the publications were extracted using the following regular expression:

```
/<electronic-resource-num>[^<]*<\/electronic-resource-num>/
```

3. For each of DOI {D}, the record was obtained in SCOPUS using the following call to the SCOPUS API (with a key {K} obtained from SCOPUS):

```
http://api.elsevier.com/content/abstract/doi/{D}?apiKey={K}&view=FULL&httpAccept=application/json
```

4. From the returned JSON response, the SCOPUS identifier (ID) of each article was parsed and stored.

```
JSON path: "abstracts-retrieval-response":"coredata":"dc:identifier"
```

Annex 2. Clustering by overlapping references

The input for this procedure (made by *in-house* developed scripts in the Ruby language) is the list of SCOPUS IDs for all the references generated as shown in Annex I.

1. For each SCOPUS ID, the list of reference cited by this article was obtained using the following call to the SCOPUS API, where {ID} is the SCOPUS ID and {K} a key obtained from SCOPUS:

```
http://api.elsevier.com/content/abstract/scopus_id/{ID}?apiKey={K}&view=FULL&httpAccept=application/json
```

   The list of references were parsed from the JSON response:

```
JSON path: "abstracts-retrieval-response":"item":"bibrecord":"tail":"bibliography":"reference"
```

   For each of those, the SCOPUS ids were parsed from the following path, and the array of SCOPUS IDs in the references was linked to the SCOPUS ID of the original article.

```
JSON path: "ref-info":"refd-itemidlist":"itemid":"$"
```

2. For all possible pairs of articles in the original list, the intersection of their arrays of SCOPUS IDs in their references was obtained. The amount of references they have in common is determined by calculating the size of this intersection, divided by the size of the smaller of the two original arrays. When this value was greater than a set threshold (i.e 0.25), a row was printed in an output file, with the following information:

```
{SCOPUS ID 1} {SCOPUS ID 2} {overlap value}
```

3. The file produced was processed by MCL[5] with the default parameters --abc -I 2) to produce clusters of articles from these pairwise comparisons.
4. This clustered file was then processed once more using the SCOPUS API to obtain the first author name, article title, journal and DOI for each SCOPUS IDs in the different clusters.

```
SCOPUS call:
http://api.elsevier.com/content/abstract/scopus_id/{ID}?apiKey={K}&view=FULL&httpAccept=application/json
JSON paths:
  First author: "abstracts-retrieval-response":"coredata":"dc:creator":"author" (first) "ce:indexed-name"
  Title: "abstracts-retrieval-response":"coredata":"dc:title"
  Journal: "abstracts-retrieval-response":"coredata":"prism:publicationName"
  DOI: "abstracts-retrieval-response":"coredata":"prism:doi"
```

5. This information was compiled into a CSV-formatted file, for import to Excel.

[5] The MCL algorithm is short for the Markov Cluster Algorithm, a fast and scalable unsupervised cluster algorithm for graphs (also known as networks) based on simulation of (stochastic) flow in graphs. It is available at http://micans.org/mcl/

Annex 3. Categorisation with keywords

The input for this procedure (made by in-house developed scripts in the Ruby language) is the list of SCOPUS IDs for all the references generated as shown in Annex I.

1. For each scopus ID, information for this article was obtained using the following call to the SCOPUS API, where {ID} is the SCOPUS id and {K} a key obtained from SCOPUS:

```
http://api.elsevier.com/content/abstract/scopus_id/{ID}?apiKey={K}&view=FULL&httpAccept=application/json
```

2. From the response, a string composed of the title, abstract and MeSH keywords was generated by concatenating (separated by spaces, and changed to downcase) the following fields:

```
JSON paths:
  Title: "abstracts-retrieval-response":"coredata":"dc:title"
  Abstract: "abstracts-retrieval-response":"coredata":"dc:description"
  MeSH keywords: "abstracts-retrieval-response":"idxterms":"mainterm":"$"
```

3. Files containing all the keywords described in section 4.1 were read and parsed. These keywords were searched in each string generated in the previous step, with simple text matching functions. Any hit assigned the corresponding SCOPUS ID to the keyword. Each dimension (Technique, Application, Aspect) was processed independently.
4. A final table was generated for Excel, using the same strategy as in Step 4 of Annex 2. The three dimensions were built as nested cells in the three first column of the Excel table (see section 4.1), and for each trio the corresponding articles were identified and printed.
5. Finally, the results were also produced in independent XML files, one file per combination of technique/application/aspect. The original records for the references in the Export file from Mendeley (Step 1 of Annex 1) were parsed and recombined accordingly for each output file, using the DOI of the hits to identify the original XML entry.

Europe Direct is a service to help you find answers to your questions about the European Union.

Freephone number (*)**:**

00 800 6 7 8 9 10 11

(*) The information given is free, as are most calls (though some operators, phone boxes or hotels may charge you).

More information on the European Union is available on the internet (http://europa.eu).

HOW TO OBTAIN EU PUBLICATIONS

Free publications:

- one copy:
  via EU Bookshop (http://bookshop.europa.eu);
- more than one copy or posters/maps:
  from the European Union's representations (http://ec.europa.eu/represent_en.htm);
  from the delegations in non-EU countries (http://eeas.europa.eu/delegations/index_en.htm);
  by contacting the Europe Direct service (http://europa.eu/europedirect/index_en.htm) or
  calling 00 800 6 7 8 9 10 11 (freephone number from anywhere in the EU) (*).

  (*) The information given is free, as are most calls (though some operators, phone boxes or hotels may charge you).

Priced publications:

- via EU Bookshop (http://bookshop.europa.eu).

KJ-NA-28618-EN-N

JRC Mission

As the science and knowledge service of the European Commission, the Joint Research Centre's mission is to support EU policies with independent evidence throughout the whole policy cycle.

EU Science Hub
ec.europa.eu/jrc

@EU_ScienceHub

EU Science Hub - Joint Research Centre

Joint Research Centre

EU Science Hub

Publications Office

doi: 10.2760/256470

ISBN 978-92-79-69019-8