eROSA
e-infrastructure Roadmap
for Open Science in Agriculture

D1.1 Bibliometric study results

Co-funded by the Horizon 2020
Framework Programme of the European Union

DELIVERABLE NUMBER	D1.1
DELIVERABLE TITLE	D1.1 Bibliometric study results
RESPONSIBLE AUTHOR	INRA

GRANT AGREEMENT N.	730988
PROJECT ACRONYM	e-ROSA
PROJECT FULL NAME	Towards an e-infrastructure Roadmap for Open Science in Agriculture
STARTING DATE (DUR.)	01/01/2017 (18 months)
ENDING DATE	30/06/2018
PROJECT WEBSITE	erosa.aginfra.eu
COORDINATOR	Odile Hologne
ADDRESS	Route de Saint-Cyr RD 10, Versailles, 78026, France
REPLY TO	odile.hologne@inra.fr
PHONE	+33 1 30 83 33 92
EU PROJECT OFFICER	Mrs. Georgia Tzenou

WORKPACKAGE N. | TITLE	WP1 | Ecosystem and Community
WORKPACKAGE LEADER	INRA
DELIVERABLE N. | TITLE	D1.1 | Bibliometric study results
RESPONSIBLE AUTHOR	Sophie Aubin (INRA)
REPLY TO	sophie.aubin@inra.fr
DOCUMENT URL	http://www.erosa.aginfra.eu/sites/erosa_deliverables/D1.1.pdf
DATE OF DELIVERY (CONTRACTUAL)	31 March 2017 (M3)
DATE OF DELIVERY (1ST SUBMISSION)	24 April 2017 (M4)
DATE OF DELIVERY (2ND SUBMISSION)	25 june 2018 (M18)
VERSION | STATUS	V4 | Final updated
NATURE	R(Report)
DISSEMINATION LEVEL	PU(Public)
AUTHORS (PARTNER)	Sophie Aubin (INRA), Madeleine Huber (INRA)

VERSION	MODIFICATION(S)	DATE	AUTHOR(S)
V1	Draft	14/03/2017	Sophie Aubin (INRA)
V2	Pre-final	28/03/2017	Sophie Aubin (INRA), Madeleine Huber (INRA)
V3	Final	24/04/2017	Sophie Aubin (INRA), Madeleine Huber (INRA)
V4	Final updated	25/06/2018	Sophie Aubin (INRA)

V4/V3 changes: The results of the bibliometric study where only accessible in the PDF report (V3) and not easy to explore. To answer the request of the stakeholders, a web site has been developed and is accessible here http://erosa-study.inra.lodex.fr/. The deliverable D.1.1 has been updated to take into account this new output with a new paragraph 2.1.3. A section on Key Findings has been added after the executive summary. These are the differences between V3 and V4.

PARTICIPANTS		CONTACT
Institut National de la Recherche Agronomique (INRA, France)		Odile Hologne Email: odile.hologne@inra.fr
Stichting Wageningen Research (Alt-DLO, The Netherlands)		Sander Janssen Email: sander.janssen@wur.nl
Agro-Know IKE (Agroknow, Greece)		Nikos Manouselis Email: nikosm@agroknow.com

ACRONYMS LIST

WoS	Web of Science™
AD	Address field in Web of Science
DE	Author Keywords field in Web of Science

EXECUTIVE SUMMARY

This document presents the identification of scientific networks & key partners, i.e. the researchers specialised in information technologies and computing and linked to agriculture.

With this network analysis, we seek to understand how the research institutions in our domain (i.e. agriculture) tackle the data challenge. In particular, this study has allowed to:

- identify key researchers and research institutions that focus on this issue and analyse to what extend they collaborate;
- link these stakeholders to specific topics that relate to the data science issue in agricultural research in order to better interpret the landscape of identified stakeholders.

This study is based on the analysis of selected publication records that are simultaneously related to the fields of data and agriculture in the broad sense of those terms.

We make no claims on comprehensiveness as the study consists in a first scoping step. Also, we do not aim to qualify excellence through this analysis as we have rather focused on the understanding of "who does what" and the network of collaborations. The analysed publications cover the period of the past decade 2005-2015 at the world scale.

This deliverable has been elaborated in the context of the e-ROSA project (Towards an e-infrastructure Roadmap for Open Science in Agriculture) under the overall Task 1.1 "Setting the stage". The latter aims in particular at mapping key stakeholders, scientific communities, infrastructures and initiatives that can support the elaboration of and benefit from an e-infrastructure roadmap for open science in the field of agriculture. Thus, a main outcome of this deliverable is the understanding of the potential relevance of the used method (i.e. bibliometric analysis) to map the stakeholder community that we seek to engage within the context of the e-ROSA project.

KEY FINDINGS

Among the key findings of the study is the growing interest of the agriculture and nutrition community for Open Science and Big data as shown by the constant increasing number of publications and diversity of contributors on data management and sharing, and related topics. In the global landscape, EU plays a leading role in providing the skills and knowledge to tackle the data challenge through key players such as INRA – particularly in knowledge engineering – and WUR – in modelling – and a rich collaborating network of countries including Spain, the Netherlands, Denmark, Germany and Greece.

In addition, there is also a strong leadership of non-European countries and organisations, especially China and the USA, with significant collaboration amongst these two countries, China having a strong focus on the Internet of Things (e.g. sensors) and the USA on information technology. The international dimension of the open data movement in the agricultural field is illustrated by the involvement of other organisations in India and Australia, as well as the FAO, which plays a central role in collaborating with both European and non-European institutions and in supporting efforts on Open Data, Linked Data and semantics.

As a result, the potential for capacity building and upscaling of required skills and knowledge to produce and use FAIR data in the agricultural field is significant at European and international levels.

ACKNOWLEDGEMENTS

The authors are grateful to Hugues Leiser and Marie-Violaine Tatry (INRA) for their valuable expertise on the methodology and conduct of a bibliometric analysis, and for their help in providing and analysing the data required for this study. They would also like to thank the e-ROSA Consortium partners that have contributed to the study as validators of the analysed corpus of publications as well as reviewers of earlier drafts of this report.

TABLE OF CONTENTS

1 INTRODUCTION

1.1 BACKGROUND

The e-ROSA project seeks to build a shared vision of a future sustainable e-infrastructure for research and education in agriculture in order to promote Open Science in this field and as such contribute to addressing related societal challenges. In order to achieve this goal, e-ROSA's first objective is to bring together the relevant scientific communities and stakeholders and engage them in the process of co-elaboration of an ambitious, practical roadmap that provides the basis for the design and implementation of such an e-infrastructure in the years to come.

The initial need for scoping and precisely mapping the e-ROSA-related stakeholder community is addressed in the context of Work Package 1 "Ecosystem & Community". In particular, the present deliverable is the outcome of the sub-task T1.1.1 "Bibliometric analysis to identify scientific networks & key partners", which aimed at providing a starting basis for the scoping of the e-ROSA community.

1.2 OBJECTIVE OF THE STUDY

This study highlights the results of a bibliometric analysis conducted at a global scale in order to identify key scientists and associated research performing organisations (e.g. public research institutes, universities, Research & Development departments of private companies) that work in the field of agricultural data sources and services.

The added value of such a methodological approach is the resulting ability to provide a detailed answer to the question "who does what?" by collecting, processing, analysing and visualising the metadata[1] of related scientific publications. The study focuses on articles that have been published in the past 10 years (i.e. during the period 2005-2015). As such, the analysis is a first attempt at delineating, mapping and describing the scientific community that the e-ROSA project seeks to engage with. It neither aims at being

[1] Metadata related to a scientific article consists of data enabling the identification and referencing of the article (e.g. authors, title, abstract, etc.)

exhaustive nor at providing an evaluation on the scientific excellence of identified stakeholders as this is not the goal of the community-building activity under e-ROSA.

The specific objectives of the analysis include:

- The identification of scientists and related collaboration networks involved in data science for agriculture in order to initiate further contact while building and engaging with the e-ROSA community throughout the project: e.g. these results provide valuable contacts in the context of the desk surveys that will be carried out under Work Package 1 in order to consolidate and reach out to the community, and in the context of the workshops organised under Work Package 2 that seek community-building and co-design of the e-ROSA Roadmap.
- The identification of specific domains related to data and computer science that are of interest to identified scientists (i.e. working on agricultural issues).
- The identification of related conferences and journals that the e-ROSA project can target in order to effectively reach out to the relevant communities involved in data science issues related to agriculture.

1.3 OVERVIEW OF THE TYPES OF OBTAINED RESULTS

The analysis has allowed to elaborate:

- a corpus of relevant publications on data production, management, use and sharing in agriculture;
- collaboration maps between stakeholders at the country and institutional levels;
- a thematic map linking institutions to a specific topic related to data science;
- the lists of persons specialised in specific topics (restricted data, i.e. for project purpose only);
- graphs illustrating community practices (i.e. journals and conferences via which scientists disseminate their results).

2 METHODOLOGY

2.1 CONSTITUTION OF THE CORPUS

2.1.1 Source choice

The Web of Science (WoS) was chosen amongst several possible sources for the following reasons: this bibliographic database has broad domain coverage, including sciences, social sciences and humanities, and is particularly relevant in the fields of agriculture, nutrition and environment, yet with some possible gaps in data science. It provides access to up-to-date bibliographic information on a wide variety of selected content and offers rich metadata, in particular the author's organisation and country, which consists of pieces of information of major interest to our study. WoS is used here via INRA's subscription.

2.1.2 Query definition

The following query was applied (01/24/2017) on the Web of Science database, using SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH indexes, with a time restriction to the 2005-2015 period, and a document type restriction to Article, Book, Book chapter, Proceedings paper and Review. Selected references were to contain both terms from the "data" domain (in green) and from the agriculture and food domains (in orange and in blue).

> TS=("information technolog*" or "information management" or "meta-information" or "e-infra*" or "big data" or "linked data" or "data linkage" or "web data" or "data reliability" or "data infrastructure*" or "data cyberinfrastructure*" or "data harmonization" or "data warehouse$" or "data management" or "data model?ing" or "data intensive computing" or "distributed computing" or "high performance computing" or "digital data" or "research data" or "open data" or "data lake" or "public dataset*" or "web collaborative" or "web based system*" or "open science" or "IoT" or "Internet of things" or "Internet technol*" or "Cloud based" or "sensor-based system*" or RDF or RDFs or OWL or XML or SPARQL or Hadoop or MapReduce or Spark or NoSQL or "MIEL system" or "conceptual graph*" or "semantic*" or "ontolog*" or NLP or "natural language process*" or "vocabular*" or "term alignment" or "term extraction" or "term recognition" or "terminological resource*" or "real-time analysis" or "knowledge engineering" or "knowledge retrieval" or "knowledge base*" or "knowledge representation" or ((knowledge or "graph-based") near/3 reasoning) or "distributed graph*" or "learning agent*" or "e-learning" or "machine learning" or "algorithm learning" or "agricultural model?ing" or "agricultural database*" OR "digital sensor*") and TS=("agricultur*" or "agronom*" or "agrifood" or "agro?environmental" or "cultural system*" or "crop system*" or "supply chain*" or "farming" or "farm$" OR cattle OR livestock) OR TS=(("agricultur*" or "agronom*" or "agrifood" or "agro?environmental" or "cultural system*" or "crop system*" or "supply chain*" or "farming" or "farm$" OR cattle OR livestock) near/3 ("information system$"))

The query resulted in 6409 publications.

2.1.3 Corpus validation

This initial corpus of 6409 publications needed to be checked and filtered out from publications irrelevant to our subject. For this purpose, it was divided in sections with decreasing confidence order:

- In sub-corpus 1 (1461 records), keywords from the above query appear either in the title or in the author keywords of the publication;
- In sub-corpus 2 (276 records), keywords appear (i) in the title/author keywords for keywords denoting "agriculture" and (ii) in the abstract for those denoting "data" (or the reverse);

- In sub-corpus 3 (1471 records), keywords appear in the abstract;
- In sub-corpus 4 (3201 records), keywords appear in the KW+ field, i.e. added by the Thomson Reuters editorial team.

Excerpts of 200 records for each of the three first sub-corpora were presented to the project experts for validation purpose. They were asked to tag each entry depending on its relevance to our study with the following tags: Y is for "Yes" (i.e. the publication is relevant), N for "No" (i.e. the publication is not relevant) and U for "Uncertain" (i.e. the validator cannot say whether the publication is relevant or not). The validation results are presented in Table 1.

	Sub-corpus 1		Sub-corpus 2		Sub-corpus 3	
	Y	Y or U	Y	Y or U	Y	Y or U
Validator 1	51%	51%	33%	33%	24%	25%
Validator 2	86%	93%	35%	39%	45%	45%
Validator 3	61%	86%	44%	66%	61%	68%

Table 1. Corpus expert validation results on 200 records

As highlighted in this table, the three validations provided mixed results individually. This can be related to the broadness of the scope of the e-ROSA project. First, the latter may have triggered a difficulty in precisely defining the query of the corpus extraction to fully reflect the e-ROSA scope and hence affected the quality of the acquired corpus. For example, the validation phase revealed that the term "supply chain" in the query brought a great number of publications unrelated to agriculture. Second, assessing the precise relevance of a specific publication was challenging in regards to such a broad scope.

On the basis of this validation phase, sub-corpus 1 was used as a reference in order to identify publications that were in scope and those out of scope within the entire corpus extracted initially. This was achieved thanks to machine learning techniques (see below).

2.1.4 Corpus delimitation

After a first cleaning of Sub-corpus 1 based on relevance of the papers (hence reducing the initial sub-corpus size from 1461 articles down to 1413), the sub-corpus was manually divided into three sets:

- 626 documents in the scope of e-ROSA,
- 647 out of scope: the latter were for instance related to the terms "owl" (as the animal, not the data format standard), "wind farm" (related to wind power), "supply chain" not relating to agriculture or food;
- 140 considered neither in scope nor out of scope (i.e. the level of confidence was not significant enough to associate these papers to one category or the other).

In order to identify the irrelevant publications of sub-corpora 2, 3, and 4, a categorisation model was trained on the binary set of 1273 documents (i.e. the first two sets described above) from sub-corpus 1. Categorisation refers to the classification of documents in predefined categories, in our case "in scope" / "out of scope". For this task we used the Luxid® Category Workbench (now known as Cogito Studio[2]) that

[2] http://www.expertsystem.com/products/cogito-studio/

performs learning and assignment of categories on the basis of terminological similarity between documents. Results are presented in Table 2 below.

	Sub-corpus 1	Sub-corpus 2		Sub-corpus 3		Sub-corpus 4		Total	
		predicted	corrected	predicted	corrected	predicted	corrected	predicted	corrected
Valid	**765**	1235	**1187**	1148	**1078**	647	**628**	3034	**3658**
Invalid	647	755	837	293	378	731	791	1779	2653

Table 2. Corpus extension. Results as predicted by the machine learning algorithm (predicted) and after human validation (corrected)

The final corpus in scope contains 3658 records of scientific publications in computer science applied to agriculture and related domains. As useful information, the metadata related to each publication mention: authors and their affiliation, the publication type, the publication date, the authors keywords, and the journal or conference they were published in.

2.2 DATA PREPARATION

To produce collaboration and thematic maps, we crossed two types of information:

- the organisations, i.e. the affiliations of authors available in the metadata of publications ("AD" WoS field);
- the topics, available as Author Keywords in the metadata ("DE" WoS field).

Some work on organisation polishing was required in order to extract the name of the organisation from the author's complete address field (AD) that contains the country, the organisation and generally one or several subdivisions (department, units, etc.). Issues such as variants, merged organisations and ambiguities (e.g. French and Moroccan INRA) were overcome during the extraction process.

Also, the journals and conferences in which scientific papers were published or presented were extracted in order to identify dissemination community practices amongst scientists. Conference names were normalised in order to regroup under a same name (i) the successive versions of an event (e.g. "PROCEEDINGS OF 2005 INTERNATIONAL CONFERENCE ON MACHINE LEARNING AND CYBERNETICS, VOLS 1-9" and "PROCEEDINGS OF 2007 INTERNATIONAL CONFERENCE ON MACHINE LEARNING AND CYBERNETICS, VOLS 1-7") and (ii) the variants in naming.

2.2.1 Topic detection

"Author Keywords" is a free text field as no controlled vocabulary or simply recommendations are proposed to authors when the latter choose the keywords at submission time for an article. This results in a great variation in keywords including morphological and lexical variants, e.g. singular vs. plural, and also in granularity, e.g. referring to very specific issues vs. broad scientific fields. A normalisation task was performed using Sphinx Lexica, a software tool that performs lexical analysis of large amounts of texts. More commonly used to process survey results, this tool is also used by INRA to prepare data for bibliometric studies.

As the topic of e-infrastructures is very broad, a single map with all stakeholders and Author Keywords together would be difficult to read. Thus, in order to facilitate the interpretation of the study results and to overcome the issue of variation in granularity of author keywords, we decided to identify specific topics

of interest. Four categories related to data had already been defined from the very beginning of the e-ROSA project:

1. big data
2. sensors
3. semantics
4. modeling

We used clustering techniques on the corpus to 1) validate these foreseen categories and 2) potentially identify additional ones. In order to perform this exploratory task, we used the clustering module of Luxid®, which provided the clustering results below. Figure 2 provides a more visual representation of the results given in Figure 1. The size of the cluster is proportional to the number of documents it contains while the width of the line between two clusters is proportional to the number of descriptors (or terms) they have in common. The clustering tool is based on the same technology, i.e. terminological similarity, as the categorisation module discussed above in Section 2.1 (Luxid® Category Workbench). In contrast, it does not rely on preexisting categories and forms categories on its own.

Description du cluster	Analyse	Documents
Non classés		22 docs
data , information , system , management , service , decision , information system , resource , analysis , process		697 docs
gene , expression , sequence , analysis , cattle , science , protein , research , sci , pathway		466 docs
knowledge , ontology , system , domain , concept , information , model , expert system , research , development		403 docs
farmer , product , information , traceability , logistics , system , safety , management , adoption , business		384 docs
model , change , area , forest , landscape , use , watershed , river , simulation , habitat		354 docs
classification , image , method , machine , algorithm , data , prediction , cow , technique , accuracy		354 docs
sensor , internet , application , network , thing , system , IoT , RDF , environment , data		339 docs
development , country , policy , welfare , economy , research , conservation , biotechnology , health , sector		243 docs
student , language , course , education , university , e-learning , word , project , science , mathematics		234 docs
energy , emission , biomass , cost , power , engine , fuel , system , gas , use		162 docs

Figure 1. Clusters on 3658 documents. 10 clusters on the basis of Author Keywords and generic term extraction - Textual view

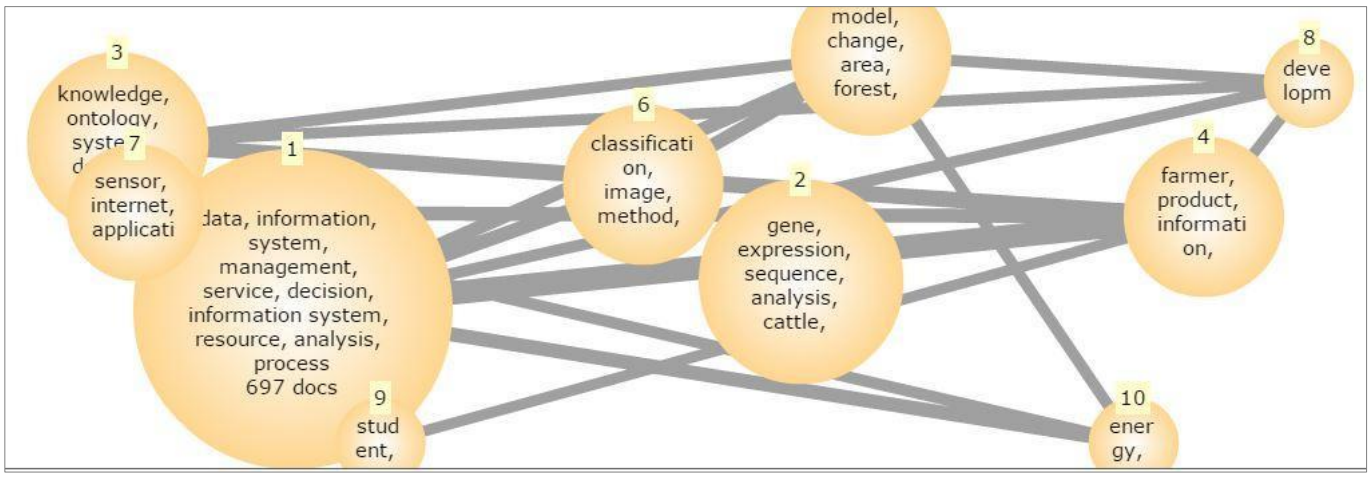

Figure 2. Clusters on 3658 documents. 10 clusters on the basis of Author Keywords and generic term extraction - Graphical view

The clustering revealed three additional topics:

1. knowledge transfer
2. information systems
3. decision support

2.2.2 Keyword organization

In order to link the studied corpus to the seven topics listed above, we associated each of the 4500 Author Keywords of the corpus with one or several topics. Project partners were again solicited for this task. All keywords appearing in at least three documents and almost all appearing in two documents were looked at.

- 949 keywords were assigned to at least one “data” topic (big data, semantics, modeling, etc.)
- 326 keywords were qualified as denoting agriculture and related fields
- 681 were qualified as useless in the context of our study

In the following, we use the term “data Author Keywords” to refer to the first list of 949 keywords, and the term “agriculture Author Keywords” to refer to the second list of 326 keywords.

2.3 PRODUCTION OF GRAPHS AND NETWORK MAPS

2.3.1 Production of graphs and network maps with CorTexT

CorTexT[3] is a collaborative platform that was developed by INRA. It consists of various modules that allow to process, analyse and structure textual data extracted from bibliographic corpora. As such, it can provide quantitative indicators (e.g. number of publications per organisation) as well as qualitative analyses via the detection and visualisation of networks (e.g. map of the relationship of co-publication between organisations). In this study, we used CorTexT to produce graphs of publication frequency of institutions and countries as well as collaboration networks at the country, institutional and authors’ levels. All maps have been generated using the CorTexT « map heterogeneous » module with the Louvain method for community detection[4].

The maps are to be read as follows:

- The node size and the line width are proportional to the number of considered objects (number of publications for an author, a country, a keyword, etc.).
- Colored circles represent the limits of the clusters, their surface being proportional to the number of nodes that they include.

For the sake of readability of the maps, thresholds were applied on data. For countries, only the 50 most publishing ones are considered. For authors and institutions, the 150 most publishing were considered.

2.3.2 Production of word clouds

In order to provide greater comprehension of the content of each topic defined in Section 2.2 above, seven word clouds have been generated in order to visualise this content in terms of Author Keywords (see Annex 1). We used Wordle[5], an online free tool that provides a graphical representation of a set of

[3] https://docs.cortext.net/

[4] https://www.quora.com/Is-there-a-simple-explanation-of-the-Louvain-Method-of-community-detection

[5] http://www.wordle.net/

words according to their frequency in a corpus. This was applied to the Authors Keywords in their original form, i.e. without any normalization. Keywords with a frequency of 1 or 2 are not represented.

2.3.3 Online publication of the results in LODEX

In addition to this report, the data were published online in an interactive portal. This allows anyone to access, query and manipulate the survey data in a graphical interface. The technology we use provides means to better understand the collected data and focus on data of particular interest. We use LODEX[6], a tool developed at Inist-Cnrs, France which is based on Linked Open Data technologies.
The data is accessible at this address: http://erosa-study.inra.lodex.fr/
When the user accesses the homepage, he discovers a set of graphs that propose different views on the data. The presented graphs correspond to almost all the graphs presented in this document. Each graph can be viewed individually by clicking on “Voir les détails”, and manipulated using combinable filters on data fields, e.g. country, date, topic, etc. The graphs then dynamically update accordingly.
The data used to produce is shown as a list of readable documents accompanied with their metadata. Those lists as well as graphs can be exported for further use or publication.
The added value of this presentation of the bibliometric study data and results is the possibility to produce focused analyses and also see and manipulate the source data.
Visitors are invited to comment and provide feedback to the bibliometric study and its presentation in the website through an online form: https://goo.gl/forms/YCzUREk3bw0hvQP93.

[6] https://github.com/Inist-CNRS/lodex

3 RESULTS

In this section, we present the results of the bibliometric study in the form of graphs and maps. In Sections 3.1 and 3.2, we focus on stakeholders, i.e. countries, institutions and persons. We then look at the data under a thematic point of view in Section 3.3. Finally, Section 3.4 provides some additional insights on the community publishing practices.

As a first result, Figure 3 shows the evolution of publication numbers between 2005 and 2015, revealing the constant dynamic of researches in computer sciences for agricultural data. Related data is presented in Annex 2, under “Publication number over years”.

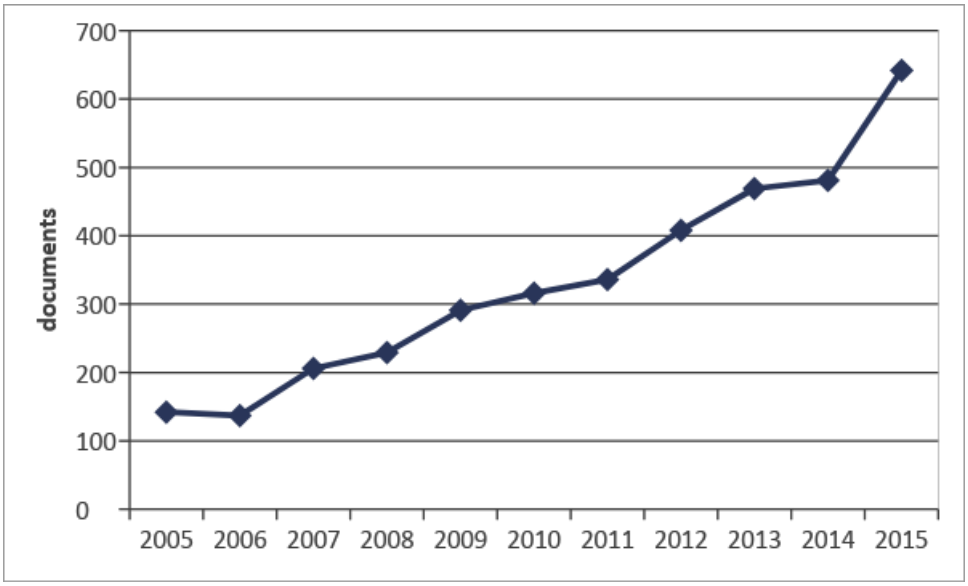

Figure 3. Temporal evolution of the 3658 publications between 2005 and 2015

3.1 COLLABORATION MAP AT COUNTRY LEVEL

China and the USA represent respectively 21% and 20% of the total amount of publications. The following are India, Germany, United Kingdom, and Australia ranging from 6.6% to 5.4%. In total, 111 countries are represented in the corpus. Related data is presented in Annex 2, under “Countries”.

Figure 4 shows groups of collaborating countries. Those represented with the same colour collaborate more with one another than with the rest of the countries. The width of the lines that connect the nodes together and the proximity between the nodes are proportional to the number of collaborations (i.e. publications) between two countries. The size of a node is proportional to the number of articles that the considered country has published in collaboration with at least one other country.

Important note: In WoS, international institutions such as FAO are affiliated to the country their headquarters are located in. This introduces a bias in some maps that we present in this study, starting with the fact that Italy (location of FAO’s headquarters) is connected to the English-speaking cluster of countries when we would expect it to be part of the green cluster that includes its direct neighboring countries. Indeed, the publications co-authored by FAO represent 18% of the total number of Italian publications.

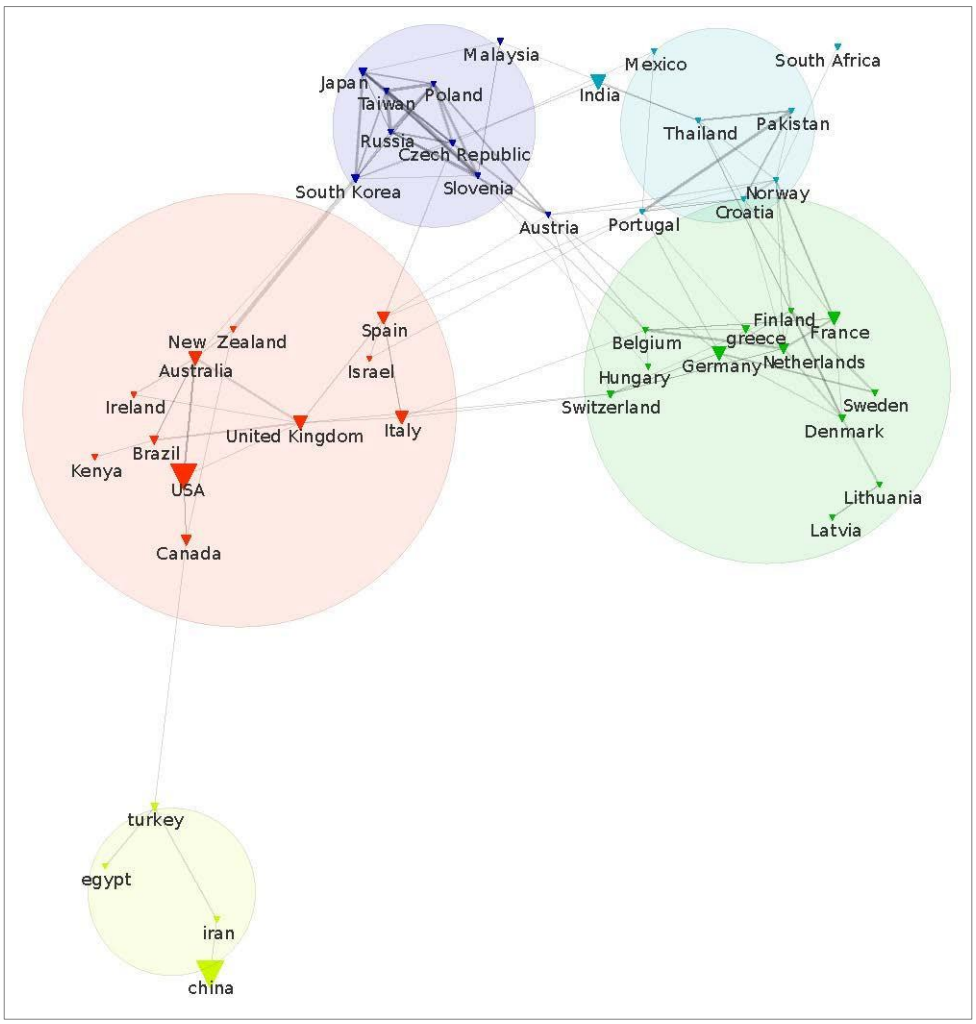

Figure 4. Collaborations between the 50 most publishing countries between 2005 and 2015

3.2 INSTITUTIONAL COLLABORATION

In order to identify key players in agricultural data production, management and use, we first built the graph in Figure 5 that shows the institutions that publish the most (related data in Annex 2 under "Institutions with more than 10 publications"). Two partners of the e-ROSA project, WUR and INRA, appear in the top 5, which in a way reinforces their legitimacy within the e-ROSA Consortium. However, this graph also demonstrates the necessity and urgency to associate non-European countries, including China, the US, India and Brazil, to support our goals. Figure 6 completes this view by revealing the institutions which repeatedly collaborate through common publications.

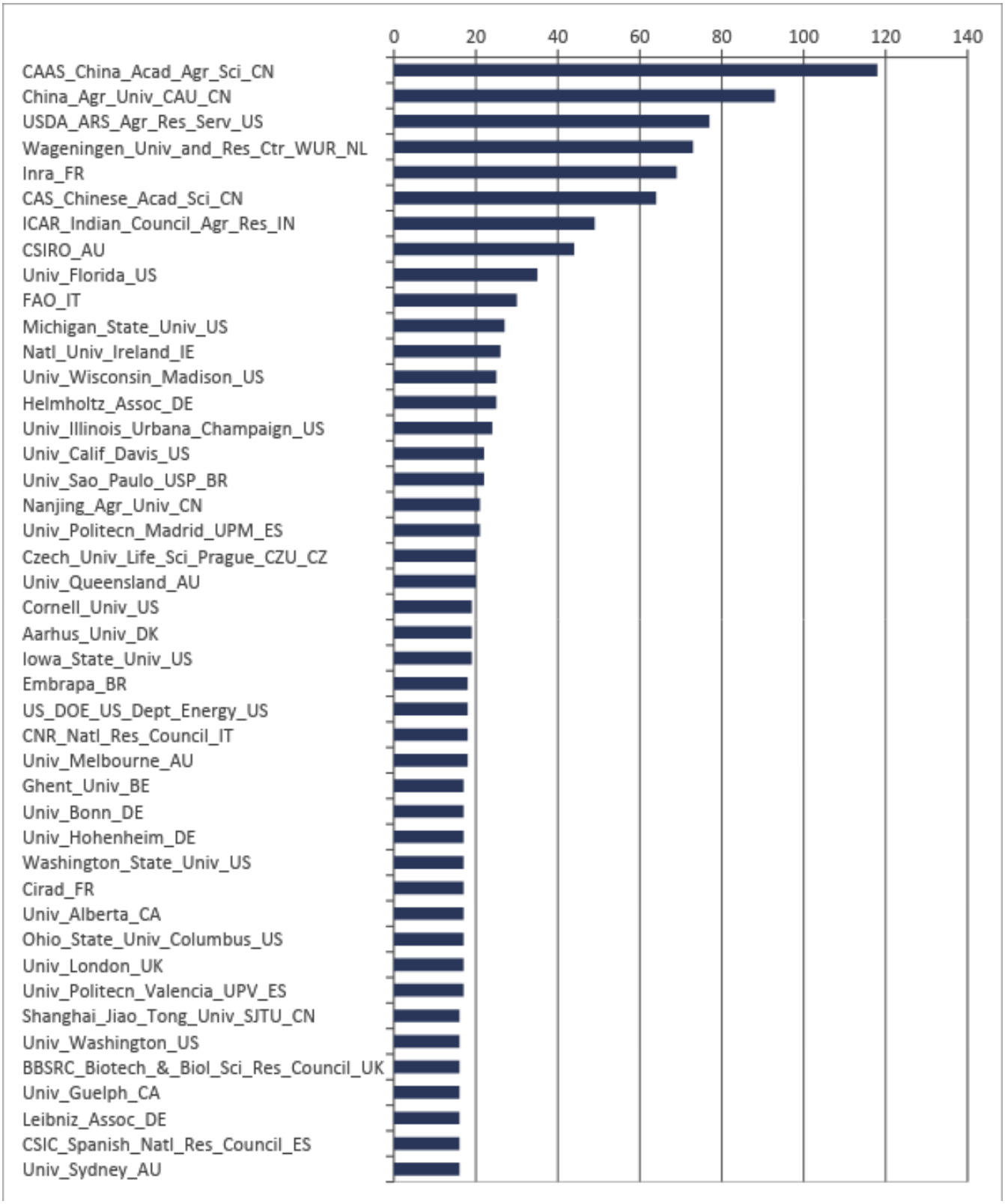

Figure 5. Most publishing institutions between (>15 publications) between 2005 and 2015

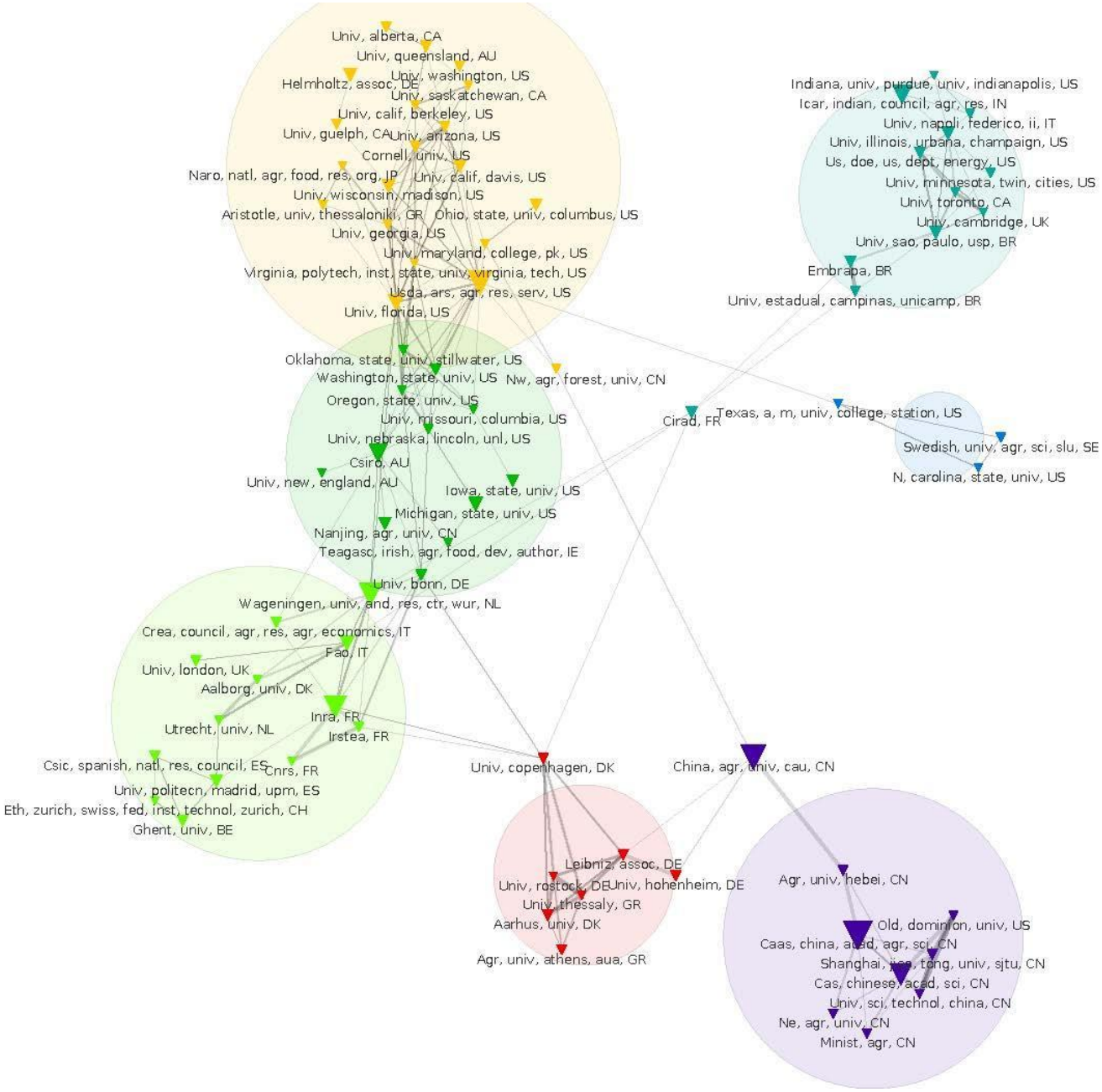

Figure 6. Collaborative networks at the institutional level (150 most publishing institutions between 2005 and 2015)

The same graphs at the authors' level have been elaborated and are considered as restricted data. The name of the most publishing authors will be of high value for the e-ROSA community-building activity.

3.3 THEMATIC ANALYSIS

3.3.1 Relationship between data issues and agricultural research fields

In Figure 7, we sought to highlight which data approaches and techniques are preferably applied by the different communities depending on the considered agricultural domain. To this end, we built the relational map of terms from the "data" set that were used to query WoS with those of the "agriculture" set (we further use the term "query keywords" to refer to the keywords used in the query described in Section 2.1.2). Data-related terms aggregate around the agriculture-related term with which they more specifically appear in the Authors Keywords.

Although already providing some interesting insights on current practices, this map could be enhanced by some preliminary keywords selection and merging. This will be done in the bibliometric study update.

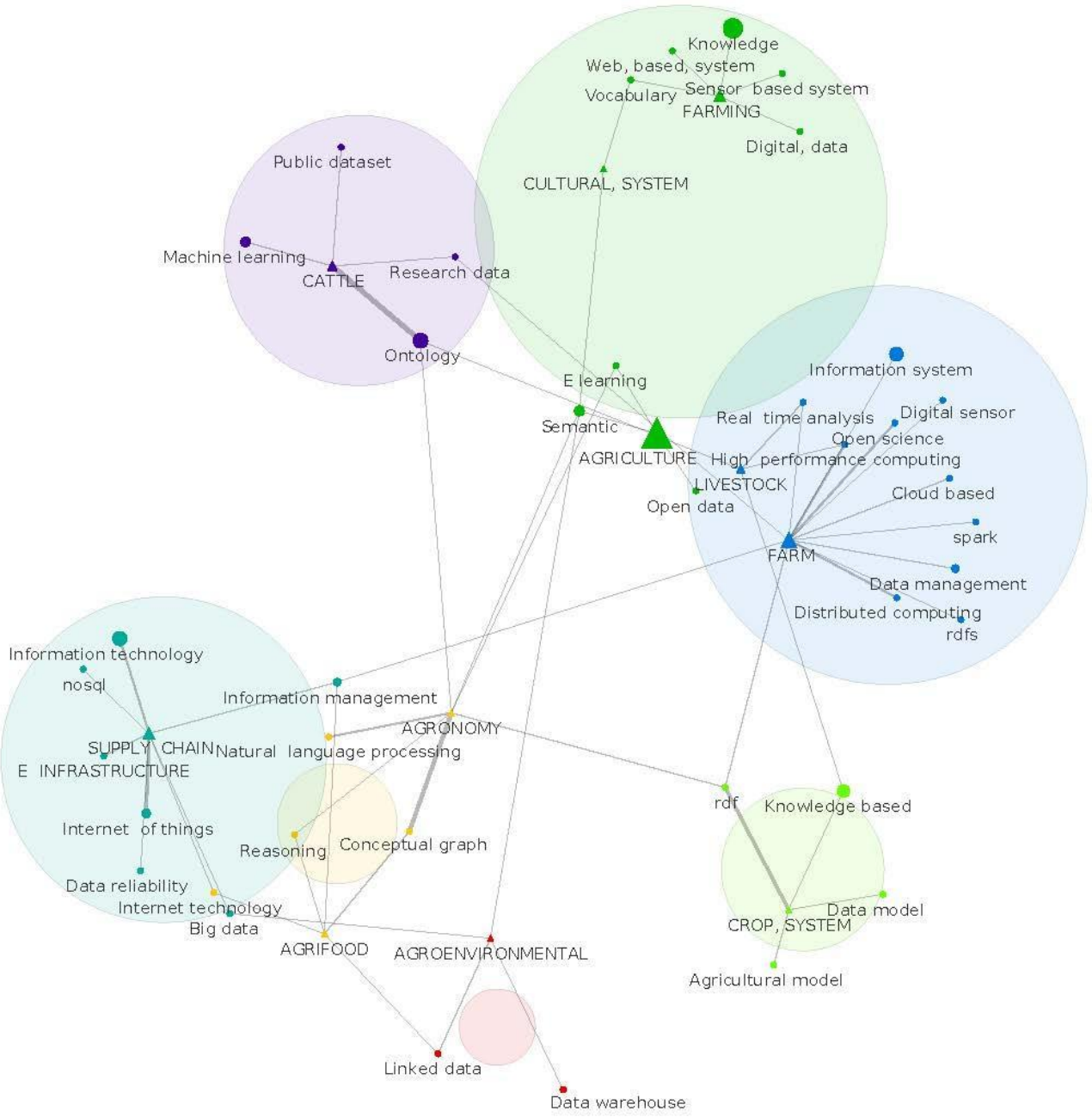

Figure 7. Relationship between "data" and "agriculture" query keywords

Three groups of clusters are especially highlighted in Figure 7:

1. The three clusters above related to the agricultural keywords "Farm", "Agriculture", "Livestock", "Cultural system", "Farming" and "Cattle": this group focuses on the farming system as a whole (however with a bias on the term "Farm" as the latter can also refer to computing farms, i.e. a data-related concept) and highlights the issue of data management and storage at this scale, in particular regarding herd management;
2. The two clusters below on the left related to "Supply Chain", "Agrifood" and "Agronomy", which focus on the technologies for efficient data management and sharing in a more significant way;
3. The cluster on "Crop system", which focuses on modelling.

3.3.2 Focus on data query keywords

3.3.2.1 *Temporal evolution of data query keywords*

Figure 8 below gives an overview of the temporal evolution of the most significant query keywords, which have been harmonised in contrast to Author Keywords. This map was build using the "Correspondence Analysis" statistical technique[7] that provides the spatial representation of a contingency table. Related data is presented in Annex 2, under "Contingency matrix of significant data keywords in query".

The size of the squares is proportional to the number of publications per year or per keyword. A square related to a specific keyword is placed near the square(s) related to the year(s) the keyword is most specific to. For instance, "machine learning", "internet of things" and "big data" became hot topics only recently around 2015. Similarly, "knowledge" appears to be a transversal concept that is not specifically related to one year or another.

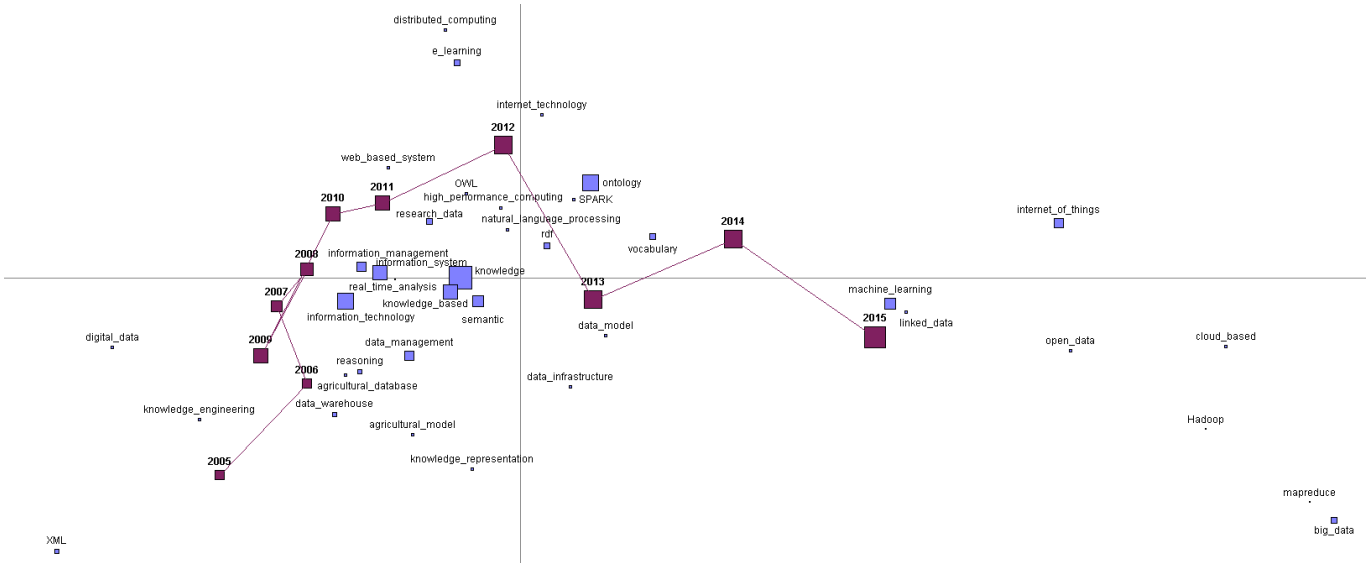

Figure 8. Temporal evolution of key data-related issues

Figure 9 provides another perspective on the data as it shows the "average age" of each query keyword in our corpus, which amounts to a year that is close to that (or those) of Figure 8. For instance, the keyword "knowledge engineering" has an average age of 2009 while the average age of "big data" is 2015, which confirms the results described in the figure above. In addition, Figure 9 shows the links between keywords in time, i.e. the width of the lines being proportional to the number of times the considered keywords are cited together. For example, "data management" will be cited in time with terms that relate to two different issues, i.e. "internet of things" on the one hand, and "machine learning" and "open science" on the other hand.

[7] Hirschfeld, H.O. (1935) "A connection between correlation and contingency", *Proc. Cambridge Philosophical Society*, 31, 520–524

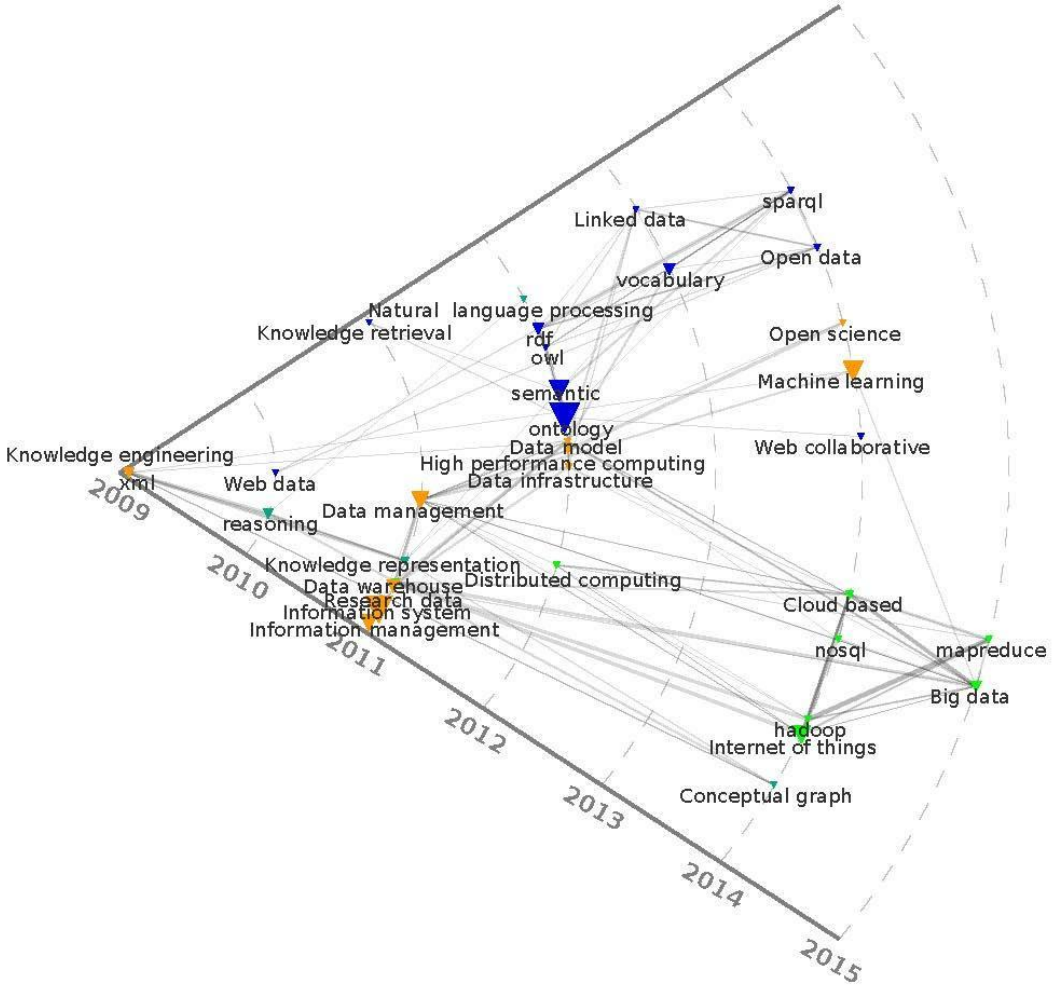

Figure 9. Historical map of data-related query keywords

3.3.2.2 *Link between data query keywords and institutions*

Figure 10 provides a deeper comprehension of the institutional landscape as institutions are linked through common subjects of interest (i.e. data-related query keywords). This graph does not necessarily show actual collaborations but also potential ones on specific subjects. In addition, it also shows connections between different fields of interest. In particular, it confirms the central role of FAO on Open Data, Linked Data and vocabularies.

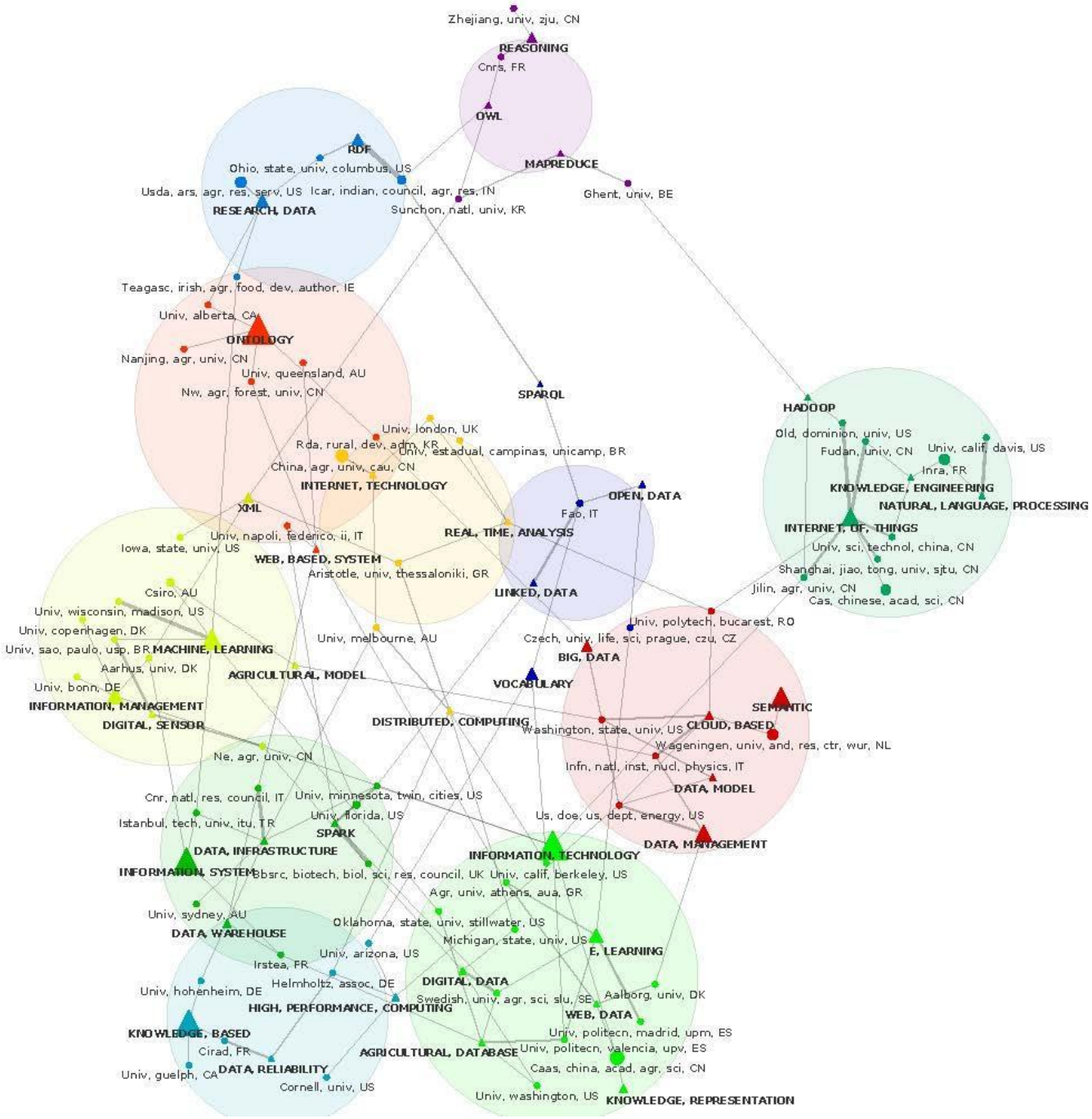

Figure 10. Clusters of institutions per data-related query keyword

3.3.3 Thematic clustering of organisations per topic

By focusing on data-specific topics of interest, sub-communities of organisations were identified. As described in Section 2.2.1 under “Topic detection”, we organized author keywords in seven categories:

1. big data
2. sensors
3. semantics
4. modeling
5. knowledge transfer
6. information systems
7. decision support

The thematic coverage of the seven topics is represented in more detail in the word clouds provided in Annex 1. Related data for the number of publications tagged with those topics is presented in Annex 2 under “Representation of topics”.

By linking the data-related author keywords with one or several topics, each topic was linked to a set of publications, and hence to related institutions that publish more specifically on the considered topic (see Figure 11 below).

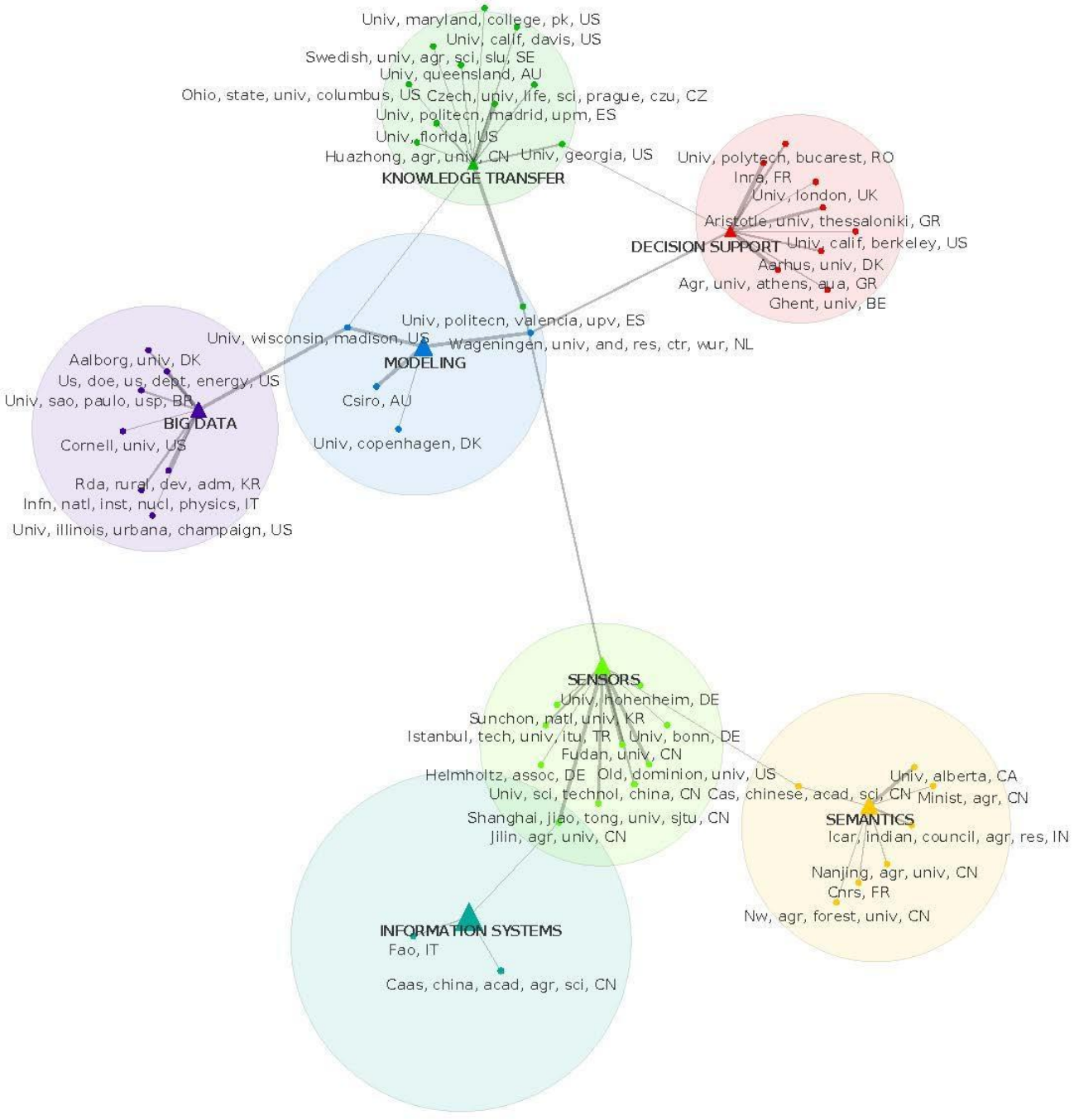

Figure 11. Network of institutions per topic

Seven additional maps were built in order to illustrate the contribution of authors to the seven topics listed in 2.2.1. As they mention individuals, they are considered as restricted data. They are accessible to project partners only.

3.4 COMMUNITY PRACTICES FOR DISSEMINATION

In this section, we have identified the most popular conferences (see Figure 12) and journals (see Figure 13) of the community in order to choose the most relevant ones that could serve future communication means of the e-ROSA project. Related data is presented in Annex 2 under “Most popular conferences” and “Most popular journals”.

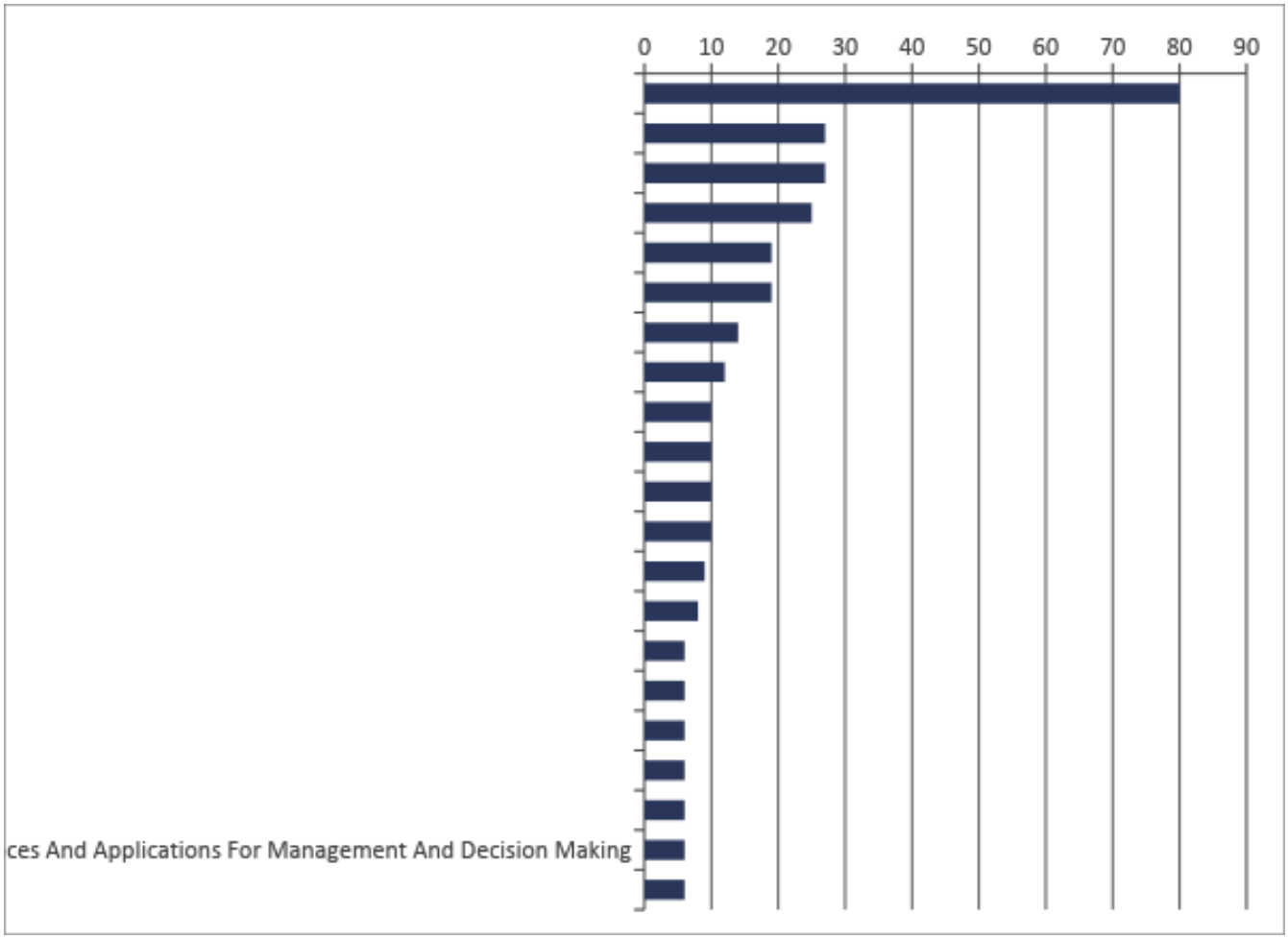

Figure 12. Most popular (>5 publications from 2005 to 2015) conferences and workshops out of a total of 913

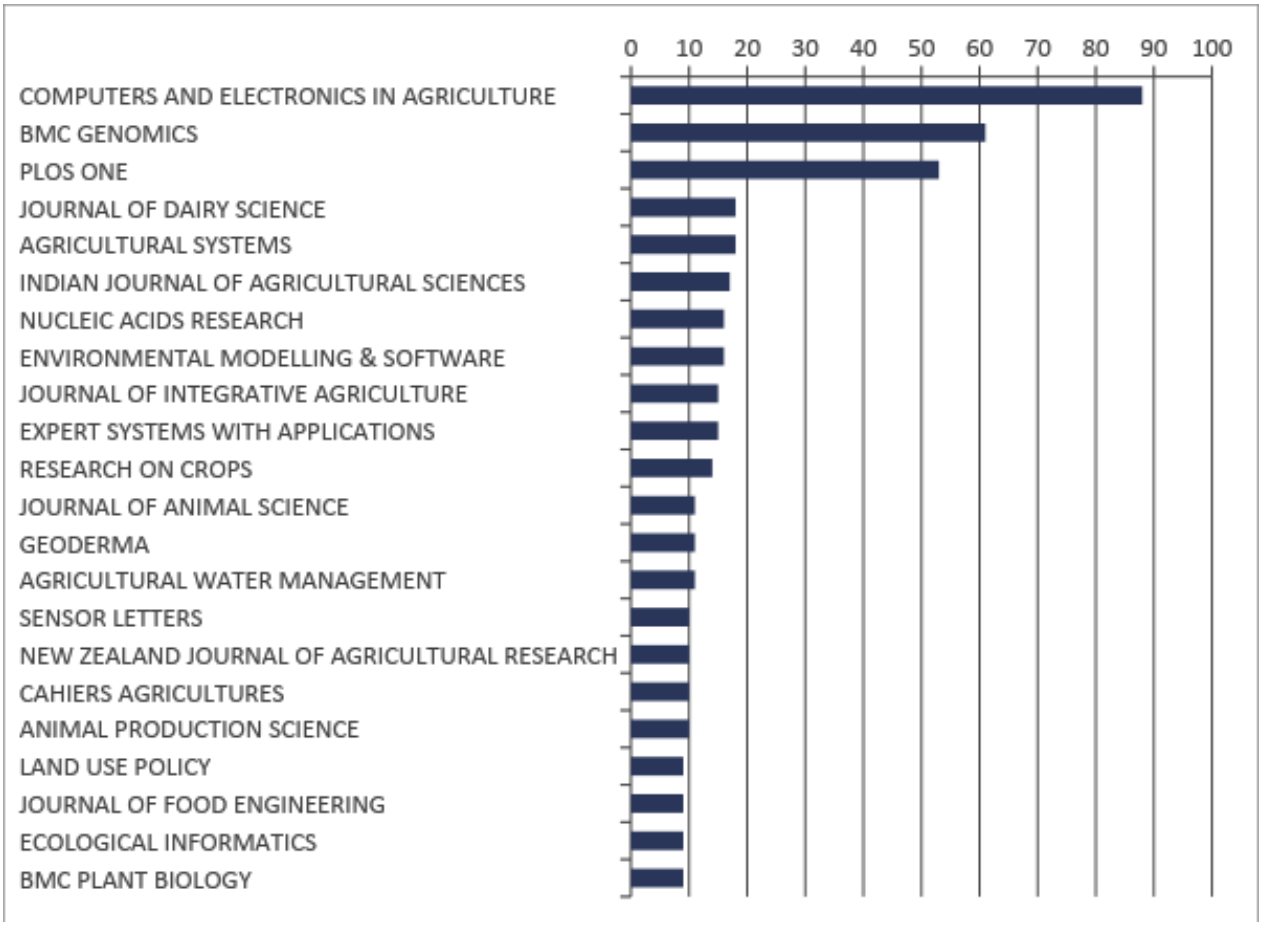

Figure 13. Most popular (>8 publications from 2005 to 2015) journals out of a total of 1058

4 DISCUSSION

4.1 STRENGTHS OF THE APPROACH

This bibliometric study has allowed to extract very detailed information and data that will be of significant value for the e-ROSA project. Indeed, acquired results have allowed to partially answer the question of *who* is currently working on agricultural data science and related issues. The identification of individual scientists and stakeholders, as well as research institutes and other organisations provides a first overview of the key experts the e-ROSA project should reach out to in order to support the co-development of its stakeholder community around the common objective of elaborating an e-infrastructure roadmap in the field of agriculture.

Furthermore, the study consists in a first attempt at describing the heterogeneous landscape of agricultural data science. As such, specific fields of interest (i.e. topics related to the theme of e-infrastructures) have been defined in order to identify sub-communities and related organisations that are specialised in these topics. Such results are of high value especially in regards to e-ROSA's objective of mapping existing initiatives and organisations that provide specific knowledge base and e-infrastructure services. Overall, this exercise of structuring and categorising the existing landscape via the thematic analysis methodology will feed into further reflections on the detailed definition of an e-infrastructure and on the various purposes the latter should serve.

4.2 DIFFICULTIES ENCOUNTERED AND LIMITATIONS

As the scope of the study is very broad, clearly delineating it was a challenging issue. In particular, defining a query that sufficiently reflects the scope in a precise manner has a significant impact on the volumetry and quality of the corpus initially extracted from WoS. As explained in Section 2.1 under "Corpus validation", a large number of publications within our initial corpus were evaluated as out of scope. Thus, refining the query may be required later on during the e-ROSA project in order to ensure greater relevance of the updated corpus in regards to the e-ROSA scope.

Also, the broadness of the e-ROSA scope has triggered difficulties in defining the topics of interest related to data and e-infrastructures (see 2.2.1 "Topic detection"). The reflection on these topics may have not yet matured enough as we are at the very beginning of the e-ROSA project. In particular, it seems that the query is focused on data management and does not reflect enough the data production issue. For instance, the bioinformatics community is underrepresented in the results of the study. In time and once e-ROSA project partners have actually reached out to scientific communities in order to understand their needs and how they can contribute to the project, relevant key issues and topics will become clearer and will be able to feed into the updating process of the bibliometric study results.

Furthermore, the stakeholders that deal with data-related issues in a scientific field such as agriculture have very diverse profiles and do not necessarily publish scientific articles that demonstrate their activity and achievements. Indeed, in many organisations, individual people that work on these issues are not only part of teams of scientists and experts but also of support teams. As the bibliometric analysis approach cannot fully capture this diversity, the obtained results describe a partial picture of the broad e-ROSA stakeholder community.

Lastly, the definition phase of the scope of the study and of the adopted approach and methodology has highlighted the importance of strongly involving project partners and experts throughout the entire analysis. Indeed, this significantly influences the quality and relevance of the elaborated methodology, the developed query and hence the extracted corpus as explained above, as well as the defined topics of

interest that feed into the thematic analysis and that highlight specialised stakeholders and organisations. In addition, certain constraints linked to the methodology developed by INRA's bibliometric experts were imposed for this first analysis (e.g. categorisation of author keywords, harmonisation of query keywords only). This triggered some inconsistencies in the methodology as a whole. For instance, associating institutions to thematic clusters of query keywords was not possible in comparison with author keywords. The updating process of the study results will seek to remove identified inconsistencies. Lastly, the outcomes achieved thanks to this study lack a practical format that allows the efficient exploitation and use of the produced data by project partners for effective outreach towards identified stakeholders. Thus, there is a need to go beyond the visualisation of results (i.e. through graphs and network maps) and explicitly link the metadata from publications in order to clearly identify the relationship between authors, institutions, countries and topics of interest. This issue will be addressed when updating the study results (see Section 5.2 below).

5 CONCLUSION

5.1 NEXT STEPS UNDER E-ROSA

This bibliometric analysis has provided a valuable preliminary inventory of key scientists and organisations active in our field of interest. The next mapping phase of the e-ROSA project will consist of an active outreach to relevant stakeholders and initiatives, including the scientists and organisations that have been identified in the bibliometric study. This will allow us to understand how they can contribute as a broad community with various interests and assets to the development of a shared vision of a future e-infrastructure for agricultural science.

As a result, e-ROSA will provide a much more advanced map of the various stakeholders relevant for the e-ROSA project and especially for its community-building activity. In particular, the precise description of identified initiatives, institutions and existing e-infrastructures, and the inputs of related stakeholders and scientific communities will be made accessible via the AgINFRA/e-ROSA portal at http://www.aginfra.eu/discover.

The actual community-building process will especially rely on the organisation of workshops targeted at the e-ROSA stakeholder community in order to provide networking opportunities. This is an essential prerequisite in order to effectively support the collaborative design of the e-infrastructure Foresight Roadmap in the context of the e-ROSA project.

ANNEX 1 – WORD CLOUDS PER DATA-RELATED TOPIC

Figure 14. Author keywords for the "Big data" topic (occurrence>2)

Figure 15. Author keywords for the "Sensors" topic (occurrence>2)

Figure 16. Author keywords for the "Semantics" topic (occurrence>2)

Figure 17. Author keywords for the "Modeling" topic (occurrence>2)

Figure 18. Author keywords for the "Knowledge transfer" topic (occurrence>2)

Figure 19. Author keywords for the "Decision support" topic (occurrence>2)

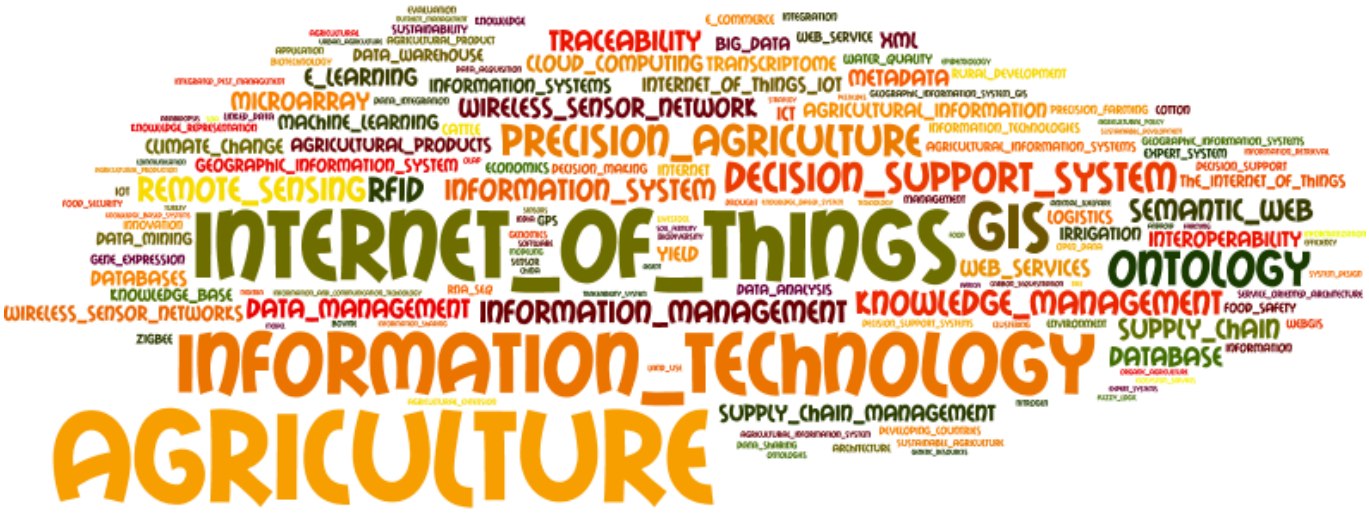

Figure 20. Author keywords for the "Information systems" topic (occurrence>2)

ANNEX 2 – DATA

Publication number over years

2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
142	137	206	229	291	316	336	408	469	481	642

Countries

Country	Documents	Part
China	767	20,97%
USA	731	19,99%
India	241	6,59%
Germany	213	5,82%
United_Kingdom	205	5,61%
Australia	196	5,36%
Italy	171	4,68%
Spain	164	4,48%
France	154	4,21%
Canada	114	3,12%
Netherlands	113	3,09%
Brazil	77	2,11%
Japan	75	2,05%
Greece	73	2,00%
South_Korea	69	1,89%
Turkey	59	1,61%
Switzerland	56	1,53%
Romania	55	1,50%
Denmark	47	1,29%
Belgium	45	1,23%
Malaysia	43	1,18%
Taiwan	43	1,18%
South_Africa	42	1,15%
Czech_Republic	41	1,12%
Sweden	39	1,07%
Iran	37	1,01%
Thailand	37	1,01%
Ireland	36	0,98%
New_Zealand	36	0,98%
Poland	36	0,98%
Norway	35	0,96%
Austria	34	0,93%
Finland	30	0,82%
Latvia	27	0,74%
Mexico	25	0,68%
Kenya	23	0,63%
Pakistan	23	0,63%
Argentina	19	0,52%
Slovakia	18	0,49%
Sri_Lanka	17	0,46%

Israel	16	0,44%
Slovenia	16	0,44%
Egypt	15	0,41%
Hungary	15	0,41%
Nigeria	15	0,41%
Russia	15	0,41%
Lithuania	14	0,38%
Croatia	13	0,36%
Portugal	13	0,36%
Colombia	12	0,33%
Indonesia	12	0,33%
Singapore	12	0,33%
Chile	9	0,25%
Ethiopia	9	0,25%
Saudi_Arabia	9	0,25%
Serbia	9	0,25%
Tanzania	9	0,25%
Philippines	8	0,22%
Bulgaria	6	0,16%
Costa_Rica	6	0,16%
Morocco	6	0,16%
Uganda	6	0,16%
Ghana	5	0,14%
Oman	5	0,14%
Tunisia	5	0,14%
Vietnam	5	0,14%
Cuba	4	0,11%
Qatar	4	0,11%
Uruguay	4	0,11%
Algeria	3	0,08%
Arab_Emirats	3	0,08%
Bangladesh	3	0,08%
Bosnia_Herzegovina	3	0,08%
Botswana	3	0,08%
Cameroon	3	0,08%
Estonia	3	0,08%
Jordan	3	0,08%
Nepal	3	0,08%
Peru	3	0,08%
Zimbabwe	3	0,08%
Benin	2	0,05%
Cambodia	2	0,05%
Iceland	2	0,05%
Jamaica	2	0,05%
Madagascar	2	0,05%
Panama	2	0,05%
Senegal	2	0,05%

Ukraine	2	0,05%
Venezuela	2	0,05%
Albania	1	0,03%
Angola	1	0,03%
Barbados	1	0,03%
Bolivia	1	0,03%
Cyprus	1	0,03%
El_Salvador	1	0,03%
Guatemala	1	0,03%
Haiti	1	0,03%
Kazakstan	1	0,03%
Lebanon	1	0,03%
Macedonia	1	0,03%
Malawi	1	0,03%
Maldives	1	0,03%
Mali	1	0,03%
Mauritius	1	0,03%
Mongolia	1	0,03%
Mozambique	1	0,03%
Myanmar	1	0,03%
Nicaragua	1	0,03%
Niger	1	0,03%
Rwanda	1	0,03%
Zambia	1	0,03%
TOTAL	3657	

Institutions with more than 10 publications

Institution	Country	Documents	Part
CAAS_China_Acad_Agr_Sci	CN	118	3,20%
China_Agr_Univ_CAU	CN	93	2,50%
USDA_ARS_Agr_Res_Serv	US	77	2,10%
Wageningen_Univ_and_Res_Ctr_WUR	NL	73	2,00%
Inra	FR	69	1,90%
CAS_Chinese_Acad_Sci	CN	64	1,80%
ICAR_Indian_Council_Agr_Res	IN	49	1,30%
CSIRO	AU	44	1,20%
Univ_Florida	US	35	1,00%
FAO	IT	30	0,80%
Michigan_State_Univ	US	27	0,70%
Natl_Univ_Ireland	IE	26	0,70%
Univ_Wisconsin_Madison	US	25	0,70%
Helmholtz_Assoc	DE	25	0,70%
Univ_Illinois_Urbana_Champaign	US	24	0,70%
Univ_Calif_Davis	US	22	0,60%
Univ_Sao_Paulo_USP	BR	22	0,60%
Nanjing_Agr_Univ	CN	21	0,60%

Univ_Politecn_Madrid_UPM	ES	21	0,60%
Czech_Univ_Life_Sci_Prague_CZU	CZ	20	0,60%
Univ_Queensland	AU	20	0,60%
Cornell_Univ	US	19	0,50%
Aarhus_Univ	DK	19	0,50%
Iowa_State_Univ	US	19	0,50%
Embrapa	BR	18	0,50%
US_DOE_US_Dept_Energy	US	18	0,50%
CNR_Natl_Res_Council	IT	18	0,50%
Univ_Melbourne	AU	18	0,50%
Ghent_Univ	BE	17	0,50%
Univ_Bonn	DE	17	0,50%
Univ_Hohenheim	DE	17	0,50%
Washington_State_Univ	US	17	0,50%
Cirad	FR	17	0,50%
Univ_Alberta	CA	17	0,50%
Ohio_State_Univ_Columbus	US	17	0,50%
Univ_London	UK	17	0,50%
Univ_Politecn_Valencia_UPV	ES	17	0,50%
Shanghai_Jiao_Tong_Univ_SJTU	CN	16	0,40%
Univ_Washington	US	16	0,40%
BBSRC_Biotech_&_Biol_Sci_Res_Council	UK	16	0,40%
Univ_Guelph	CA	16	0,40%
Leibniz_Assoc	DE	16	0,40%
CSIC_Spanish_Natl_Res_Council	ES	16	0,40%
Univ_Sydney	AU	16	0,40%
Univ_Arizona	US	15	0,40%
Zhejiang_Univ_ZJU	CN	15	0,40%
Agr_Univ_Hebei	CN	15	0,40%
Sunchon_Natl_Univ	KR	15	0,40%
Agr_Univ_Athens_AUA	GR	15	0,40%
Minist_Agr	CN	15	0,40%
Univ_Nebraska_Lincoln_UNL	US	14	0,40%
Univ_Estadual_Campinas_UNICAMP	BR	14	0,40%
RDA_Rural_Dev_Adm	KR	14	0,40%
Swedish_Univ_Agr_Sci_SLU	SE	14	0,40%
CREA_Council_Agr_Res_&_Agr_Economics	IT	14	0,40%
Irstea	FR	14	0,40%
Univ_Minnesota_Twin_Cities	US	14	0,40%
Univ_Copenhagen	DK	14	0,40%
NE_Agr_Univ	CN	14	0,40%
Univ_Calif_Berkeley	US	14	0,40%
Univ_Polytech_Bucarest	RO	14	0,40%
Univ_Sci_&_Technol_China	CN	14	0,40%
Oklahoma_State_Univ_Stillwater	US	13	0,40%
AAFC_Agr_&_Agri_Food_Canada	CA	13	0,40%
Aristotle_Univ_Thessaloniki	GR	13	0,40%

Jilin_Agr_Univ	CN	13	0,40%
TEAGASC_Irish_Agr_&_Food_Dev_Author	IE	13	0,40%
INFN_Natl_Inst_Nucl_Physics	IT	12	0,30%
Istanbul_Tech_Univ_ITU	TR	12	0,30%
NW_Agr_&_Forest_Univ	CN	12	0,30%
Univ_Maryland_College_Pk	US	12	0,30%
Fudan_Univ	CN	12	0,30%
Huazhong_Agr_Univ	CN	12	0,30%
Texas_A&M_Univ_College_Station	US	12	0,30%
Univ_Georgia	US	12	0,30%
Univ_Napoli_Federico_II	IT	12	0,30%
Oregon_State_Univ	US	11	0,30%
Univ_Saskatchewan	CA	11	0,30%
Univ_Toronto	CA	11	0,30%
Univ_New_England	AU	11	0,30%
Univ_Missouri_Columbia	US	11	0,30%
Utrecht_Univ	NL	11	0,30%
Aalborg_Univ	DK	11	0,30%
Univ_Alcala_UAH	ES	11	0,30%
Wuhan_Univ	CN	11	0,30%
CNRS	FR	11	0,30%
Old_Dominion_Univ	US	11	0,30%
Mississippi_State_Univ	US	10	0,30%
Sojo_Univ	JP	10	0,30%
Virginia_Polytech_Inst_&_State_Univ_Virginia_Tech	US	10	0,30%
ETH_Zurich_Swiss_Fed_Inst_Technol_Zurich	CH	10	0,30%
NARO_Natl_Agr_&_Food_Res_Org	JP	10	0,30%
SW_Univ	CN	10	0,30%
Indiana_Univ_Purdue_Univ_Indianapolis	US	10	0,30%
Univ_Edinburgh	UK	10	0,30%
N_Carolina_State_Univ	US	10	0,30%
Univ_Cambridge	UK	10	0,30%
Univ_Rostock	DE	10	0,30%
Univ_Thessaly	GR	10	0,30%

Contingency matrix of significant data keywords in query

Data keyword in query /year	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	TOTAL
knowledge	**39**	**46**	**60**	**63**	**78**	**96**	**95**	**99**	**126**	**114**	**141**	**957**
ontology	**12**	**16**	**25**	**27**	**39**	**52**	**55**	**79**	**77**	**92**	**101**	**575**
information_technology	**32**	**23**	**46**	**43**	**52**	**41**	**63**	**59**	**59**	**47**	**60**	**525**
information_system	**20**	**20**	**36**	**43**	**50**	**49**	**39**	**57**	**51**	**48**	**59**	**472**

knowledge_based	22	17	23	32	35	50	34	44	54	47	64	422
semantic	12	9	13	21	26	21	23	29	37	29	42	262
machine_learning	5	6	7	10	5	11	14	27	39	49	86	259
internet_of_things	0	0	0	2	2	4	15	26	30	51	79	209
data_management	12	6	16	10	20	15	15	19	26	18	25	182
information_management	7	7	7	15	23	26	17	18	20	16	22	178
e_learning	2	1	5	10	7	12	16	17	13	15	12	110
research_data	2	4	5	10	11	10	11	10	9	14	13	99
vocabulary	2	5	3	6	3	5	9	12	13	13	20	91
rdf	4	1	5	8	5	6	3	12	15	13	12	84
big_data	0	0	0	0	0	0	0	2	14	16	47	79
XML	10	6	4	5	12	2	11	4	6	5	0	65
reasoning	4	4	5	2	7	10	3	6	9	5	8	63
data_warehouse	4	1	7	4	2	4	6	2	6	2	7	45
knowledge_representation	3	3	3	4	4	2	3	1	6	8	6	43
high_performance_computing	1	3	4	2	2	3	3	7	2	5	7	39
cloud_based	0	0	0	0	1	0	0	2	4	13	15	35
data_model	1	1	3	0	3	2	4	4	5	2	9	34
internet_technology	1	0	2	3	1	3	5	5	4	4	6	34
open_data	0	1	0	0	0	1	0	3	9	7	12	33
SPARK	0	2	3	0	2	2	1	6	8	4	4	32
data_infrastructure	1	1	0	2	5	2	3	2	3	4	7	30
distributed_computing	2	0	1	1	1	3	7	6	2	3	4	30
digital_data	2	0	3	2	6	6	2	2	4	1	1	29
linked_data	0	1	0	1	1	1	1	4	6	3	10	28
agricultural_model	2	1	1	1	4	3	2	2	5	2	4	27
OWL	1	0	3	0	3	1	2	5	3	3	3	24
web_based_system	0	2	3	2	0	4	4	1	2	3	3	24

natural_language_processing	**2**	**1**	**2**	**1**	**0**	**1**	**1**	**6**	**3**	**1**	**4**	**22**
agricultural_database	**1**	**3**	**0**	**0**	**3**	**2**	**2**	**3**	**4**	**1**	**2**	**21**
knowledge_engineering	**2**	**2**	**1**	**4**	**2**	**3**	**0**	**2**	**0**	**2**	**2**	**20**
real_time_analysis	**0**	**1**	**1**	**1**	**2**	**1**	**3**	**1**	**3**	**1**	**2**	**16**
Hadoop	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**4**	**4**	**5**	**13**
mapreduce	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**1**	**2**	**7**	**11**
digital_sensor	**2**	**0**	**1**	**1**	**0**	**1**	**0**	**1**	**3**	**0**	**0**	**9**
SPARQL	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**1**	**1**	**3**	**2**	**8**
web_data	**0**	**0**	**0**	**1**	**3**	**1**	**0**	**1**	**1**	**0**	**1**	**8**
data_reliability	**1**	**0**	**0**	**0**	**1**	**0**	**1**	**2**	**1**	**1**	**0**	**7**
sensor_based_system	**0**	**1**	**0**	**0**	**0**	**2**	**2**	**0**	**1**	**1**	**0**	**7**
noSQL	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**2**	**2**	**2**	**6**
data_harmonization	**0**	**0**	**0**	**0**	**1**	**0**	**1**	**0**	**2**	**1**	**1**	**6**
open_science	**0**	**0**	**0**	**0**	**1**	**1**	**0**	**1**	**0**	**1**	**2**	**6**
knowledge_retrieval	**0**	**0**	**0**	**0**	**0**	**2**	**1**	**0**	**0**	**2**	**0**	**5**
conceptual_graph	**0**	**0**	**0**	**0**	**2**	**0**	**0**	**0**	**0**	**2**	**1**	**5**
public_dataset	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**1**	**0**	**3**	**5**
meta_information	**1**	**0**	**0**	**0**	**1**	**0**	**1**	**0**	**0**	**0**	**0**	**3**
graph_based	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**2**	**2**
learning_agent	**0**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**2**
e_infrastructure	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**1**
RDFs	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**1**
term_extraction	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**1**
web_collaborative	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**1**
TOTAL	**212**	**195**	**298**	**338**	**427**	**461**	**480**	**591**	**696**	**681**	**926**	**5305**

Representation of topics

Topics	Documents	Part
information systems	2358	64,50%
modeling	1025	28,00%
semantics	911	24,90%
sensors	886	24,20%
big data	787	21,50%
decision support	372	10,20%
knowledge transfer	320	8,80%
uncategorized	44	1,20%
TOTAL	3658	

Most popular conferences (> 3 documents)

Name	Documents
Computer And Computing Technologies In Agriculture	80
International Conference On Agro-geoinformatics	27
Metadata And Semantics Research	27
Ieee International Geoscience And Remote Sensing Symposium	25
International Conference On Computing In High Energy And Nuclear Physics	19
International Conference On Education And New Learning Technologies	19
Progress Of Information Technology In Agriculture	14
Inted: International Technology, Education And Development Conference	12
Computational Science And Its Applications	10
Efita/wcca '	10
Information Technology Applications In Industry	10
International Conference On Computing For Sustainable Global Development	10
International Congress On Modelling And Simulation	9
International Conference On Challenges In Environmental Science And Computer Engineering	8
Actual Tasks On Agricultural Engineering	6
Asee Annual Conference	6
International Conference On Agriculture Engineering	6
International Conference On Information And Communication Technologies In Agriculture, Food And Environment	6
Isprs Congress, Technical Commission	6
International Congress On Modelling And Simulation: Advances And Applications For Management And Decision Making: Advances And Applications For Management And Decision Making	6
Rural Development	6
Advanced Science Letters	5
Agrarian Perspectives : Global Agribusiness And The Rural Economy	5
Geoconference On Informatics, Geoinformatics And Remote Sensing	5

International Conference Of Education, Research And Innovation	5
International Conference On E-learning	5
International Conference On Industrial And Information Systems	5
International Conference On Mechatronics, Electronic, Industrial And Control Engineering	5
International Conference On Wireless Communications, Networking And Mobile Computing	5
International Horticultural Congress On Science And Horticulture For People	5
International Workshop On Database And Expert Systems Applications	5
Journal Of Animal Science	5
Metadata And Semantics	5
Precision Agriculture	5
Remote Sensing For Agriculture, Ecosystems, And Hydrology	5
Research For Rural Development	5
Winter Simulation Conference	5
World Multi-conference On Systemics, Cybernetics And Informatics	5
World Imacs Congress And ModsimInternational Congress On Modelling And Simulation: Interfacing Modelling And Simulation With Mathematical And Computational Sciences	5
Wuhan International Conference On E-business	5
Applied Science, Materials Science And Information Technologies In Industry	4
Bmc Bioinformatics	4
Changing European Farming Systems For A Better Future: New Visions For Rural Areas	4
Congreso Iberico De Agroingenieria Y Ciencias Horticolas: Innovar Y Producir Para El Futuro. Innovating And Producing For The Future	4
European Conference On Knowledge Management	4
Frontiers Of Manufacturing And Design Science	4
International Conference On Applied Business Research	4
Ieee International Conference On Big Data	4
Ieee International Conference On Systems, Man, And Cybernetics	4
Iiai International Conference On Advanced Applied Informatics	4
International Conference On Advances In Computing, Communications And Informatics	4
International Conference On Informationization, Automation And Electrification In Agriculture	4
International Conference On Public Administration	4
International Scientific Conference: Engineering For Rural Development	4
International Technology, Education And Development Conference	4
International Congress On Modelling And Simulation: Land, Water And Environmental Management: Integrated Systems For Sustainability	4
Rethinking Education By Leveraging The Elearning Pillar Of The Digital Agenda For Europe!	4
World Academy Of Science, Engineering And Technology	4
Agricultural Engineering: Land - Technik Ageng : Solutions For Intelligent And Sustainable Farming	3
Animal Welfare	3
Annual Srii Global Conference	3
Applied Informatics And Communication	3
Bulletin Of The University Of Agricultural Science And Veterinary Medicine	3
Earth Resources And Environmental Remote Sensing/gis Applications	3
Ecological Modelling	3
Engineering Solutions For Manufacturing Processes	3
Enviroinfo And Ict For Sustainability	3

Ieee International Conference On Computational Intelligence And Computing Research	3
Ieee-npss Real Time Conference	3
Iiai International Congress On Advanced Applied Informatics	3
International Conference On Banana And Plantain In Africa: Harnessing International Partnerships To Increase Research Impact	3
International Conference On Computational Science	3
International Conference On Computer Technology And Development	3
International Conference On Economic, Business Management And Education Innovation	3
International Conference On Education And Sports Education	3
International Conference On Geoinformatics	3
International Conference On Industrial Control And Electronics Engineering	3
International Conference On Industrial Engineering And Engineering Management	3
International Conference On Instrumentation & Measurement, Computer, Communication And Control	3
International Conference On Intellectual Capital, Knowledge Management And Organisational Learning	3
International Conference On Machine Learning And Applications	3
International Conference On Social Sciences And Society	3
International Forum On Information Technology And Applications	3
International Workshop On Information And Electronics Engineering	3
Journal Of Food Composition And Analysis	3
Nuclear Instruments & Methods In Physics Research Section A-accelerators Spectrometers Detectors And Associated Equipment	3
Rural Environment, Education, Personality	3
Science And Information Conference	3
Tri-state Dairy Nutrition Conference	3
World Automation Congress	3
World Congress On Intelligent Control And Automation	3

Most popular journals (>3 documents)

Name	Documents
COMPUTERS AND ELECTRONICS IN AGRICULTURE	88
BMC GENOMICS	61
PLOS ONE	53
JOURNAL OF DAIRY SCIENCE	18
AGRICULTURAL SYSTEMS	18
INDIAN JOURNAL OF AGRICULTURAL SCIENCES	17
NUCLEIC ACIDS RESEARCH	16
ENVIRONMENTAL MODELLING & SOFTWARE	16
JOURNAL OF INTEGRATIVE AGRICULTURE	15
EXPERT SYSTEMS WITH APPLICATIONS	15
RESEARCH ON CROPS	14
JOURNAL OF ANIMAL SCIENCE	11
GEODERMA	11
AGRICULTURAL WATER MANAGEMENT	11
SENSOR LETTERS	10
NEW ZEALAND JOURNAL OF AGRICULTURAL RESEARCH	10
CAHIERS AGRICULTURES	10

ANIMAL PRODUCTION SCIENCE	10
LAND USE POLICY	9
JOURNAL OF FOOD ENGINEERING	9
ECOLOGICAL INFORMATICS	9
BMC PLANT BIOLOGY	9
TRANSACTIONS OF THE ASABE	8
REMOTE SENSING OF ENVIRONMENT	8
PRECISION AGRICULTURE	8
JOURNAL OF FOOD AGRICULTURE & ENVIRONMENT	8
JOURNAL OF AGRICULTURAL & ENVIRONMENTAL ETHICS	8
IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS	8
ECOLOGICAL MODELLING	8
CURRENT SCIENCE	8
BIOSYSTEMS ENGINEERING	8
ANIMAL GENETICS	8
ANIMAL	8
REMOTE SENSING	7
LEGUME RESEARCH	7
JOURNAL OF ENVIRONMENTAL MANAGEMENT	7
ENVIRONMENTAL MONITORING AND ASSESSMENT	7
BRITISH FOOD JOURNAL	7
AGRONOMY FOR SUSTAINABLE DEVELOPMENT	7
WATER RESOURCES RESEARCH	6
WATER RESOURCES MANAGEMENT	6
OUTLOOK ON AGRICULTURE	6
LIVESTOCK SCIENCE	6
JOURNAL OF VETERINARY MEDICAL EDUCATION	6
JOURNAL OF SOIL AND WATER CONSERVATION	6
JOURNAL OF RURAL STUDIES	6
JOURNAL OF ENVIRONMENTAL PROTECTION AND ECOLOGY	6
INTERNATIONAL JOURNAL OF REMOTE SENSING	6
INTERNATIONAL JOURNAL OF PRODUCTION RESEARCH	6
FUNCTIONAL & INTEGRATIVE GENOMICS	6
ELECTRONIC LIBRARY	6
CROP PROTECTION	6
COMMUNICATIONS IN SOIL SCIENCE AND PLANT ANALYSIS	6
BMC BIOINFORMATICS	6
AFRICAN JOURNAL OF AGRICULTURAL RESEARCH	6
VETERINARY IMMUNOLOGY AND IMMUNOPATHOLOGY	5
SPANISH JOURNAL OF AGRICULTURAL RESEARCH	5
RENEWABLE & SUSTAINABLE ENERGY REVIEWS	5
PREVENTIVE VETERINARY MEDICINE	5
PHYSICS AND CHEMISTRY OF THE EARTH	5
MOLECULAR BIOLOGY REPORTS	5
JOURNAL OF FOOD COMPOSITION AND ANALYSIS	5
JOURNAL OF CLEANER PRODUCTION	5
INTERNATIONAL JOURNAL OF PRODUCTION ECONOMICS	5
INTERNATIONAL JOURNAL OF ENVIRONMENTAL RESEARCH AND PUBLIC HEALTH	5

INTELLIGENT AUTOMATION AND SOFT COMPUTING	5
INDIAN JOURNAL OF TRADITIONAL KNOWLEDGE	5
GEOFORUM	5
FRONTIERS IN PLANT SCIENCE	5
ENVIRONMENTAL MANAGEMENT	5
CIENCIA RURAL	5
BMC GENETICS	5
AQUACULTURE	5
APPLIED SOFT COMPUTING	5
AGRICULTURE ECOSYSTEMS & ENVIRONMENT	5
AGRICULTURAL ECONOMICS	5
THERIOGENOLOGY	4
THEORETICAL AND APPLIED GENETICS	4
SOCIOLOGIA RURALIS	4
SENSORS	4
SCIENCE OF THE TOTAL ENVIRONMENT	4
RANGELAND JOURNAL	4
POULTRY SCIENCE	4
PLANT PHYSIOLOGY	4
PARASITES & VECTORS	4
KNOWLEDGE-BASED SYSTEMS	4
JOURNAL OF PLANT NUTRITION AND SOIL SCIENCE	4
JOURNAL OF PEASANT STUDIES	4
JOURNAL OF ENVIRONMENTAL QUALITY	4
JOURNAL OF APPLIED REMOTE SENSING	4
JOURNAL OF AGRICULTURAL SCIENCE	4
INTERNATIONAL JOURNAL OF ENGINEERING EDUCATION	4
INTERNATIONAL JOURNAL OF DISTRIBUTED SENSOR NETWORKS	4
INTERNATIONAL FOOD AND AGRIBUSINESS MANAGEMENT REVIEW	4
INFORMATION DEVELOPMENT	4
INDIAN JOURNAL OF ANIMAL SCIENCES	4
IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING	4
GLOBAL ENVIRONMENTAL CHANGE-HUMAN AND POLICY DIMENSIONS	4
GENE	4
FOOD CONTROL	4
EUROPEAN JOURNAL OF AGRONOMY	4
COMPUTERS IN INDUSTRY	4
CLIMATIC CHANGE	4
BIOMASS & BIOENERGY	4
BIODIVERSITY AND CONSERVATION	4
ARCHIV FUR TIERZUCHT-ARCHIVES OF ANIMAL BREEDING	4
APPLIED GEOGRAPHY	4
APPLIED ENGINEERING IN AGRICULTURE	4
APPLIED ANIMAL BEHAVIOUR SCIENCE	4
ANIMAL WELFARE	4
AGRICULTURE AND HUMAN VALUES	4
AGRICULTURAL SCIENCES IN CHINA	4

ADVANCES IN MODELING AGRICULTURAL SYSTEMS	4
WATER SA	3
WATER ENVIRONMENT RESEARCH	3
VETERINARY PARASITOLOGY	3
TRANSBOUNDARY AND EMERGING DISEASES	3
TIERAERZTLICHE PRAXIS AUSGABE GROSSTIERE NUTZTIERE	3
TECHNOLOGICAL AND ECONOMIC DEVELOPMENT OF ECONOMY	3
STUDIES IN INFORMATICS AND CONTROL	3
SOIL SCIENCE SOCIETY OF AMERICA JOURNAL	3
SMALL RUMINANT RESEARCH	3
SEMANTIC WEB	3
SCIENTOMETRICS	3
SCIENCE TECHNOLOGY & HUMAN VALUES	3
REMOTE SENSING LETTERS	3
PROCEEDINGS OF THE ROYAL SOCIETY B-BIOLOGICAL SCIENCES	3
PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA	3
PLANT BIOTECHNOLOGY JOURNAL	3
PHYSIOLOGICAL GENOMICS	3
PHOTOGRAMMETRIC ENGINEERING AND REMOTE SENSING	3
PERSONAL AND UBIQUITOUS COMPUTING	3
NUTRIENT CYCLING IN AGROECOSYSTEMS	3
NUCLEAR INSTRUMENTS & METHODS IN PHYSICS RESEARCH SECTION A-ACCELERATORS SPECTROMETERS DETECTORS AND ASSOCIATED EQUIPMENT	3
MOLECULAR BREEDING	3
MATHEMATICAL AND COMPUTER MODELLING	3
LIBRARY HI TECH	3
JOURNAL OF THE FACULTY OF AGRICULTURE KYUSHU UNIVERSITY	3
JOURNAL OF SUSTAINABLE AGRICULTURE	3
JOURNAL OF SENSORS	3
JOURNAL OF MANAGEMENT INFORMATION SYSTEMS	3
JOURNAL OF INDO-EUROPEAN STUDIES	3
JOURNAL OF HYDROLOGY	3
JOURNAL OF APPLIED ECOLOGY	3
JOURNAL OF ANIMAL AND PLANT SCIENCES	3
JOURNAL OF AGRICULTURAL SCIENCE AND TECHNOLOGY	3
JOURNAL FUR VERBRAUCHERSCHUTZ UND LEBENSMITTELSICHERHEIT-JOURNAL OF CONSUMER PROTECTION AND FOOD SAFETY	3
ISPRS JOURNAL OF PHOTOGRAMMETRY AND REMOTE SENSING	3
INTERNATIONAL SUGAR JOURNAL	3
INTERNATIONAL JOURNAL OF MOLECULAR SCIENCES	3
INTERNATIONAL JOURNAL OF LIFE CYCLE ASSESSMENT	3
INTERNATIONAL JOURNAL OF INFORMATION MANAGEMENT	3
INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE	3
INFORMATION-AN INTERNATIONAL INTERDISCIPLINARY JOURNAL	3
IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING	3
IEEE LATIN AMERICA TRANSACTIONS	3
HYDROLOGY AND EARTH SYSTEM SCIENCES	3

HYDROGEOLOGY JOURNAL	3
HORTTECHNOLOGY	3
GENOMICS	3
GENOME BIOLOGY	3
GENETICS AND MOLECULAR RESEARCH	3
FRONTIERS IN GENETICS	3
FOODBORNE PATHOGENS AND DISEASE	3
ENVIRONMENTAL SCIENCE & TECHNOLOGY	3
ENVIRONMENTAL SCIENCE & POLICY	3
E-AGRICULTURE AND RURAL DEVELOPMENT: GLOBAL INNOVATIONS AND FUTURE PROSPECTS	3
DECISION SUPPORT SYSTEMS	3
CUSTOS E AGRONEGOCIO ON LINE	3
CURRENT OPINION IN ENVIRONMENTAL SUSTAINABILITY	3
CONCURRENCY AND COMPUTATION-PRACTICE & EXPERIENCE	3
COMPUTING AND INFORMATICS	3
COMPARATIVE IMMUNOLOGY MICROBIOLOGY AND INFECTIOUS DISEASES	3
CATTLE PRACTICE	3
ASIA LIFE SCIENCES	3
AMERICAN JOURNAL OF AGRICULTURAL ECONOMICS	3
AGRONOMY JOURNAL	3
AGRICULTURAL ECONOMICS-ZEMEDELSKA EKONOMIKA	3
AGRICULTURAL AND FOREST METEOROLOGY	3
AGRICULTURAL AND ENVIRONMENTAL INFORMATICS, GOVERNANCE AND MANAGEMENT: EMERGING RESEARCH APPLICATIONS	3
AGRARFORSCHUNG SCHWEIZ	3

REFERENCES

1. Web of Science (WoS): https://www.webofknowledge.com
2. Cogito (Luxid) Studio: http://www.expertsystem.com/products/cogito-studio/
3. CorTexT Manager https://docs.cortext.net/
4. Correspondence analysis: Hirschfeld, H.O. (1935) "A connection between correlation and contingency", Proc. Cambridge Philosophical Society, 31, 520–524
5. AgINFRA/e-ROSA community map : http://www.aginfra.eu/discover