Design of an EMIDA database on European research institutions and their major publication topics

(Version 1 - April 2009)

Jean De Rycke

EMIDA WP2 leader

Institut National de la Recherche Agronomique,
IASP, Centre de Tours, 37380 Nouzilly, France
Tel : 33 (0)2 47 42 75 49
E-mail : jean.derycke@tours.inra.fr

TABLE OF CONTENTS

EMIDA ERA-NET is funded by the European Commission's Seventh Framework Programme, Project No. 219235

1. Objective of the database

To map research production in the field of animal health in Europe during the recent years (from 2004 onward), according to scientific topics, research organisations, and number of scientific papers and patents. The focus is on animal infectious and parasitic diseases, the chief concern of EMIDA.

2. Principles of database construction

Through data extraction from ISI Web of Knowledge [SM], to build up an Excel spreadsheet displaying the main following fields:

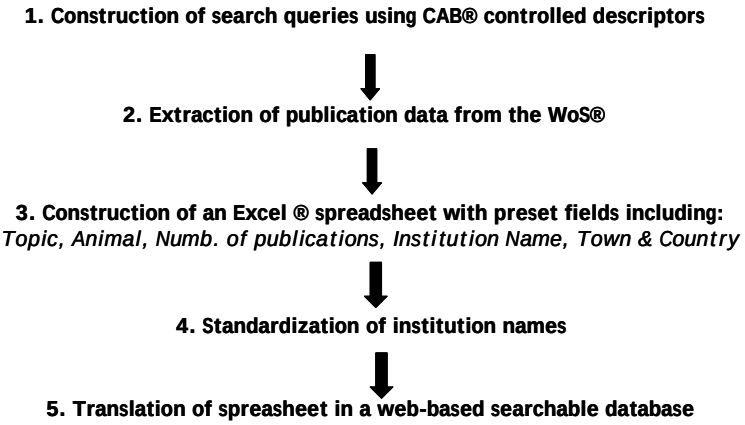
- Name of disease or pathogen
- Search query formulation
- Name of research organisation, City, Country
- Number of source records for each theme and research organization

A detailed description of the source data bases, their specific features and functionalities, can be found on the website of the corresponding platforms, run by Thomson Scientific: Thomson Reuters - ISI Web of Knowledge [SM]

3. Web-based bibliometric resources and compared potentials

Two complementary source platforms run by the ISI Web of Knowledge [SM] were used: the **CAB Abstracts ®**, specialized in agronomic research including animal health, and the **Web of Science ®** (WoS) link covering all disciplines. The advantage of CAB Abstracts ® over the WoS ® is that each source document is allocated (i) one or more standard classification codes (termed "cabicodes"), indicating the broad subject areas of the paper and (ii) several controlled descriptors (including names of pathogens, diseases and host animals) that are present in a controlled thesaurus. These features lend themselves to normalized procedures of sorting and statistical analysis of source documents, on the basis on topics. However, compared to the WoS, the CAB lacks a function that is essential for the mapping of institutions: sorting on authors' addresses, more specifically on organization names, town of location, and countries. This function is present in the WoS platform. Therefore an interplay between the two platforms was necessary.

Overall procedure



4. Construction of search queries covering the entire field of EMIDA

GENERAL DESCRIPTION

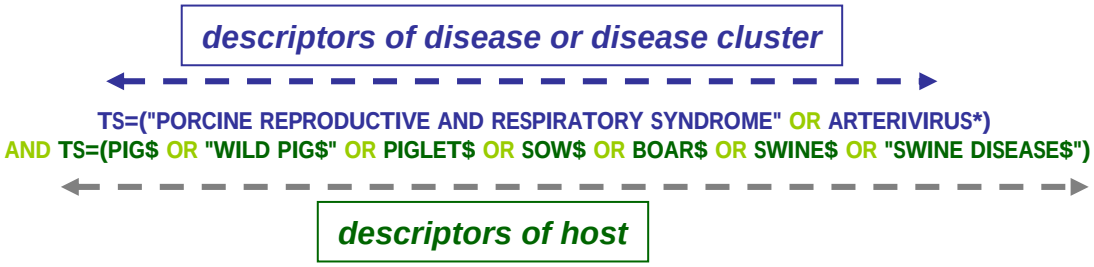
A specific challenge of the present bibliometric analysis was to construct a complete set of search queries that would cover in an exhaustive manner, and without major omissions, the research domain of EMIDA, i.e. animal infectious diseases. A particularity of this research domain is that it can be satisfactorily described by the names of diseases and of causative pathogenic agents, provided that an exhaustive list of these descriptors can be generated, then prioritized.

One way to achieve this goal could have been to exploit existing lists of diseases and of causative pathogenic agents made available by specialized expert groups. But it appears that more exhaustive lists of all the natural descriptors present in the literature could be produced in a systematic manner, then prioritised, using specific functionalities of the CAB Abstracts ® platform. Very briefly, these are the main steps leading to the formulation of search queries using descriptors controllable in the specific CAB thesaurus .

(i) Extraction of all the documents, published since 2004, identified by at least one of the three **broad area qualifiers** (called "cabicodes" or CC) that specifically describe the field of infectious and parasitic diseases of animals:
- **LL821** for “Prion, Viral, Bacterial and Fungal Pathogens of Animals”;
- **LL822** for “Protozoan, Helminth, Mollusc and Arthropod Parasites
- **LL823** for “Veterinary Pests, Vectors and Intermediate Hosts”

(ii) Identification and statistical analysis of all the **specific descriptors of diseases and pathogens** automatically generated by the platform. In the same way, identification of all possible descriptors for host animals in each food animal group.

(iii) Grouping of descriptors by a specialist of animal infectious diseases on the one hand by **"disease cluster"** (describing one disease or several closely related diseases) and on the other hand by "animal category"; then formulation of all the search queries basically composed of the two above sets of terms linked by Boolean operators and using appropriate wild cards, such as in the example represented in the diagram below:



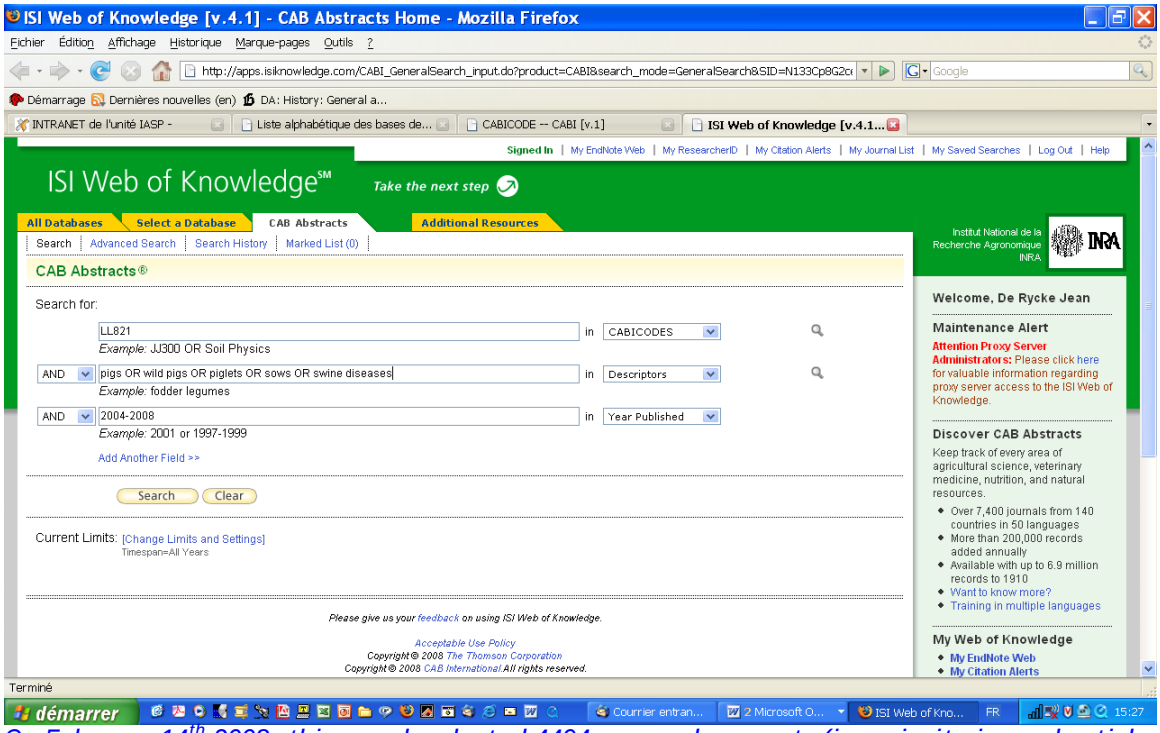
Adaptations to this general search scheme was necessary to collect data on a number of **transversal topics** such as "vaccines and vaccination", "quantitative epidemiology of infectious diseases" etc…Such searches were made feasible due to the existence of further specific "cabicodes", used in combination with other broad descriptors. The more useful cabicodes (CC) were the following:

- **LL650:** Animal Immunology; *
- **HH600:** Host Resistance and Immunity;
- **LL886**: Diagnosis of animal diseases;
- **YY700:** Pathogens, Parasites and Infectious Diseases (Wild Animals);
- **HH405**: Pesticides and Drugs: Control.

STEP BY STEP DEMONSTRATION OF DESCRIPTORS SELECTION

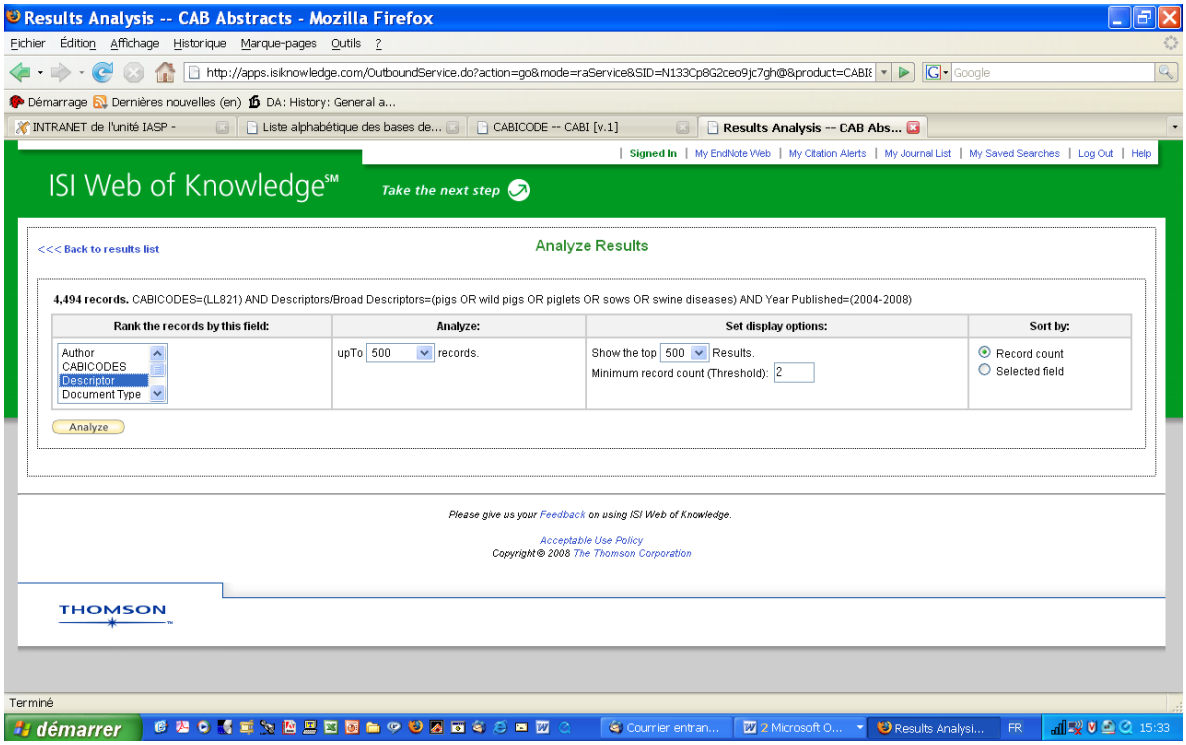
To display all possible descriptors of diseases and pathogens for a specified animal category the search query must combine the CC (Cabicode) describing a research field with thesaurus descriptors of this category of production animals"

For instance; the search query to display all possible "disease & pathogen" descriptors of bacterial, viral and prion diseases in pigs should combine the CC=LL821 with animal species CAB thesaurus controlled descriptors "pigs", "wild pigs", "piglets", "sows", "swine diseases". The result of a typical search is illustrated below:

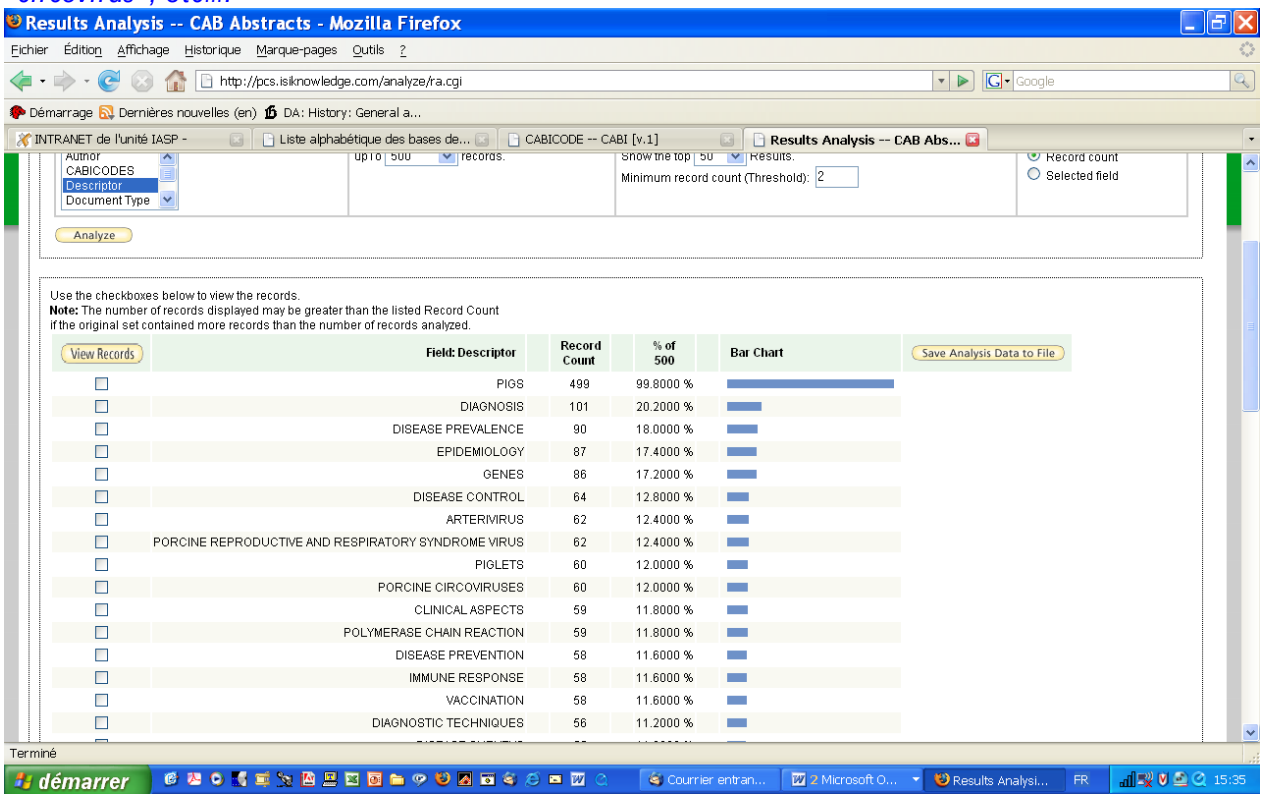


On February 14th 2008, this search selected 4494 source documents (in majority journal articles) since 2004.

As can be seen of the left column, descriptors can be sorted according to their frequency in a second step.....



...which leads to the final display below comprising up to 500 descriptors organized in tabulated format. The window shows the very first descriptors of the list, among which only those describing specific diseases and pathogens are selected. Here the most frequent descriptors referring to a pathogen or a disease are "arterivirus" and, "porcine reproductive and respiratory syndrome", which in fact points to the same disease/pathogen entity. The next most frequent descriptor is "circovirus", etc....



To select all disease/pathogen descriptors relevant to the research field searched, the complete table of descriptors is directly copied to an Excel spread sheet. From this list; the animal disease specialist (1) selects the specific descriptors of diseases and pathogens, then (2) groups together those descriptors pertaining to a same disease entity.

The selection and reorganization of controlled descriptors is based on the expertise of the animal health specialist and should be scientifically sound and accountable. Whereas this procedure of descriptors selection secures exhaustiveness and adequate prioritization, the grouping of descriptors (pointing to a same entity) may be discussed in some instances. The total traceability of the procedure (search queries are included in the final EMIDA data base) lends itself to corrections and further improvements, if deemed necessary. The reference manual used to check the relevance of disease/pathogen grouping is the Merck Veterinary Manual*, a free comprehensive electronic reference for animal diseases information.*

Once selected and reorganized the list of descriptors appears as shown (for pigs diseases covered by CC:LL821)

PORCINE REPRODUCTIVE AND RESPIRATORY SYNDROME, ARTERIVIRUS, PORCINE REPRODUCTIVE AND RESPIRATORY SYNDROME VIRUS

SALMONELLA, SALMONELLA TYPHIMURIUM, SALMONELLOSIS, SALMONELLA CHOLERAESUIS, SALMONELLA ENTERITIDIS

MYCOPLASMA HYOPNEUMONIAE, MYCOPLASMA

PORCINE CIRCOVIRUS, PORCINE CIRCOVIRUSES

SWINE FEVER, SWINE FEVER VIRUS, CLASSICAL SWINE FEVER VIRUS, AFRICAN SWINE FEVER VIRUS, AFRICAN SWINE FEVER

ACTINOBACILLUS PLEUROPNEUMONIAE, PLEUROPNEUMONIA

STREPTOCOCCUS SUIS

FOOT AND MOUTH DISEASE, APHTHOVIRUS, FOOT-AND-MOUTH DISEASE VIRUS

PASTEURELLA MULTOCIDA, ATROPHIC RHINITIS

AUJESZKY VIRUS, AUJESZKY'S DISEASE, SUID HERPESVIRUS 1, PSEUDORABIES VIRUS, VARICELLOVIRUS

ESCHERICHIA COLI

BRACHYSPIRA HYODYSENTERIAE, SWINE DYSENTERY, BRACHYSPIRA

SWINE INFLUENZAVIRUS, AVIAN INFLUENZA, AVIAN INFLUENZAVIRUS, INFLUENZA A, INFLUENZAVIRUS A

HAEMOPHILUS PARASUIS, GLASSERS DISEASES

PORCINE PARVOVIRUS, INFECTIOUS INFERTILITY

HEPATITIS E VIRUS, HEPATITIS E

BORDETELLA BRONCHISEPTICA

CAMPYLOBACTER COLI, CAMPYLOBACTER, CAMPYLOBACTER JEJUNI, CAMPYLOBACTERIOSIS

TRANSMISSIBLE GASTROENTERITIS VIRUS, WASTING DISEASE, PORCINE CORONAVIRUS

LEPTOSPIROSIS, LEPTOSPIRA, LEPTOSPIRA INTERROGANS

LAWSONIA (BACTERIA), PROLIFERATIVE ENTEROPATHY, PROLIFERATIVE ILEITIS, LAWSONIA INTRACELLULARIS

YERSINIA ENTEROCOLITICA

ROTAVIRUS

STAPHYLOCOCCUS AUREUS, STREPTOCOCCUS, MASTITIS, AGALACTIA

BRUCELLOSIS, BRUCELLA, BRUCELLA SUIS

CLOSTRIDIUM PERFRINGENS, CLOSTRIDIUM

TUBERCULOSIS, MYCOBACTERIUM AVIUM, MYCOBACTERIUM BOVIS

ERYSIPELOTHRIX RHUSIOPATHIAE

MYCOTOXICOSES, FUSARIUM, FUMONISINS, TIAMULIN

JAPANESE ENCEPHALITIS VIRUS

ENTEROCOCCUS FAECIUM, ENTEROCOCCUS, ENTEROCOCCUS FAECALIS

STAPHYLOCOCCUS HYICUS

TAENIA SOLIUM, CYSTICERCOSIS, TAENIASIS, TAENIA SAGINATA, ECHINOCOCCUS GRANULOSUS, NEUROCYSTICERCOSIS, ECHINOCOCCOSIS, TAENIA, ECHINOCOCCUS, CYSTICERCI, EUCESTODA, CESTODE INFECTIONS, METACESTODES, CYSTIC ECHINOCOCCOSIS

ASCARIS SUUM, ASCARIASIS, ASCARIS, ASCARIS LUMBRICOIDES

TOXOPLASMA GONDII, TOXOPLASMOSIS

TRICHINOSIS, TRICHINELLA SPIRALIS, TRICHINELLA, TRICHINELLA NATIVA, TRICHINELLA PSEUDOSPIRALIS

COCCIDIOSIS, ISOSPORA SUIS, EIMERIA, COCCIDIA, ISOSPORIASIS, TACHYZOITES, ISOSPORA, EIMERIA DEBLIECKI, EIMERIA PERMINUTA, EIMERIA SUIS, ENCEPHALITOZOON INTESTINALIS

CRYPTOSPORIDIUM, CRYPTOSPORIDIOSIS

TRICHURIS SUIS, TRICHURIS, TRICHURIASIS

SARCOPTES SCABIEI, ECTOPARASITOSES, SCABIES, ECTOPARASITES, PHTHIRAPTERA, DIPTERA, CTENOCEPHALIDES FELIS, DEMODEX, DERMATITIS

OESOPHAGOSTOMUM, OESOPHAGOSTOMUM DENTATUM, METASTRONGYLUS, STRONGYLOIDES RANSOMI, STRONGYLIDAE, HYOSTRONGYLUS, HYOSTRONGYLUS RUBIDUS

TOXOPLASMA , CRYPTOSPORIDIUM PARVUM, GIARDIASIS, GIARDIA, GIARDIA DUODENALIS

TICKBORNE DISEASES, RHIPICEPHALUS SANGUINEUS, METASTIGMATA, VECTOR-BORNE DISEASES, CULEX TRITAENIORHYNCHUS, BABESIA, BABESIOSIS, ANAPLASMA , EPERYTHROZOON, EPERYTHROZOON SUIS

SARCOCYSTIS, SARCOCYSTIS MIESCHERIANA

SCHISTOSOMA JAPONICUM

BALANTIDIUM COLI

LUNGWORMS

With minor modifications, the procedure is run for each CC (LL821, LL822 and LL823) and for each food animal category (pigs, ruminants, poultry, horses, rabbits, bees, fish and shellfish). All the sets of controlled descriptors are now supposed to cover in a satisfactory manner the totality of the diseases and pathogenic agents that can be searched and retrieved in the scientific literature of the domain.

These species-specific sets of descriptors can now be used as topics for further search, using different data bases, particularly WoS. One the main asset of this procedure is that all these descriptors are present in the CAB thesaurus.

5. Running search queries with controlled descriptors on WoS platform

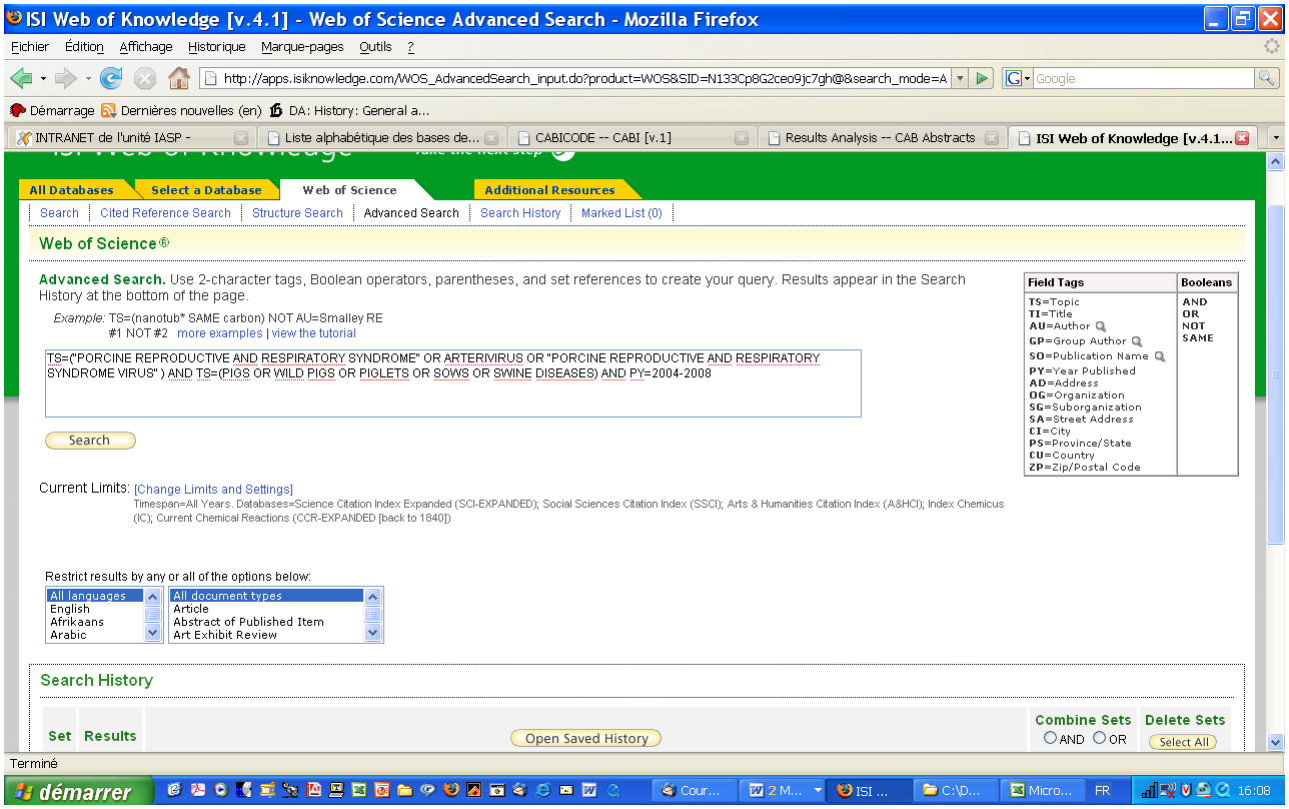
Research institutions having produced publications on each searched scientific topics since 2004 can now be identified using WoS. In contrast to the CAB; the WoS platform can sort and analyse

references using addresses, which include name of countries and of research organisations,. this for of all the authors of the source documents sorted. We have decided to include all the documents produced in 2008 at the date of the searches; so as not to exclude relevant newcomers.

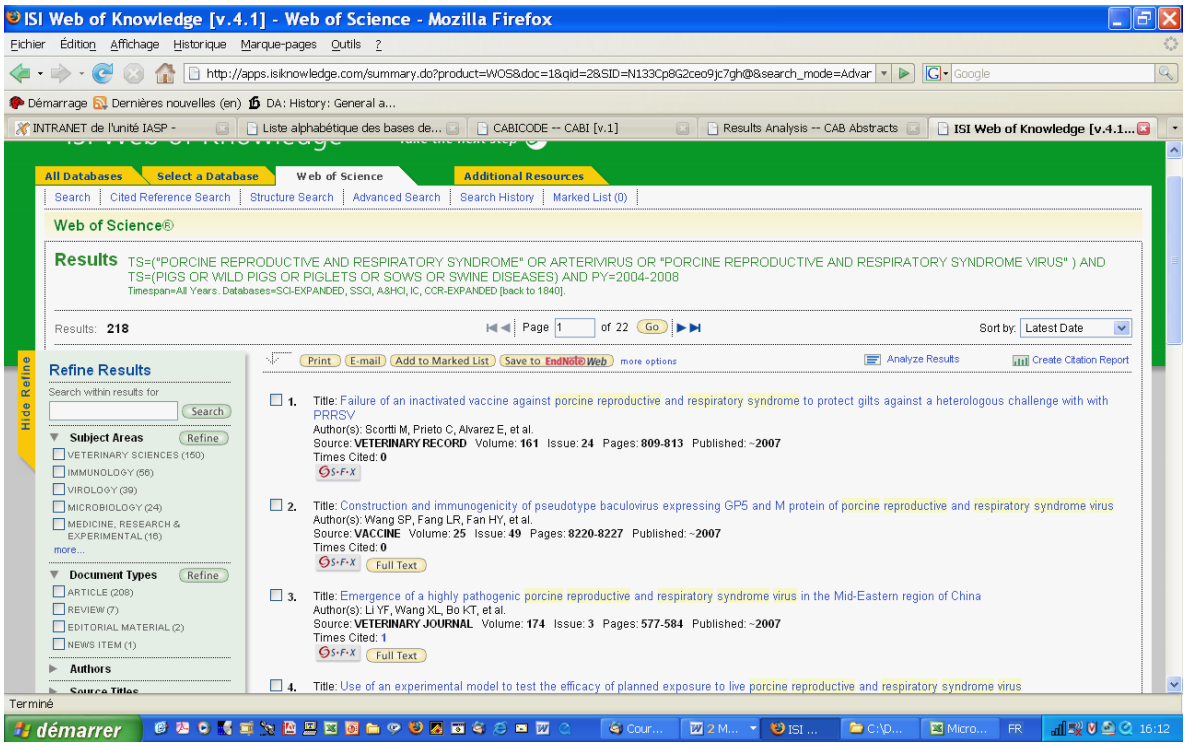
For example, the search for references on the porcine reproductive and respiratory syndrome of pigs is instructed as follows

TS=("PORCINE REPRODUCTIVE AND RESPIRATORY SYNDROME" OR ARTERIVIRUS OR "PORCINE REPRODUCTIVE AND RESPIRATORY SYNDROME VIRUS") AND TS=(PIGS OR WILD PIGS OR PIGLETS OR SOWS OR SWINE DISEASES) AND PY=2004-2008

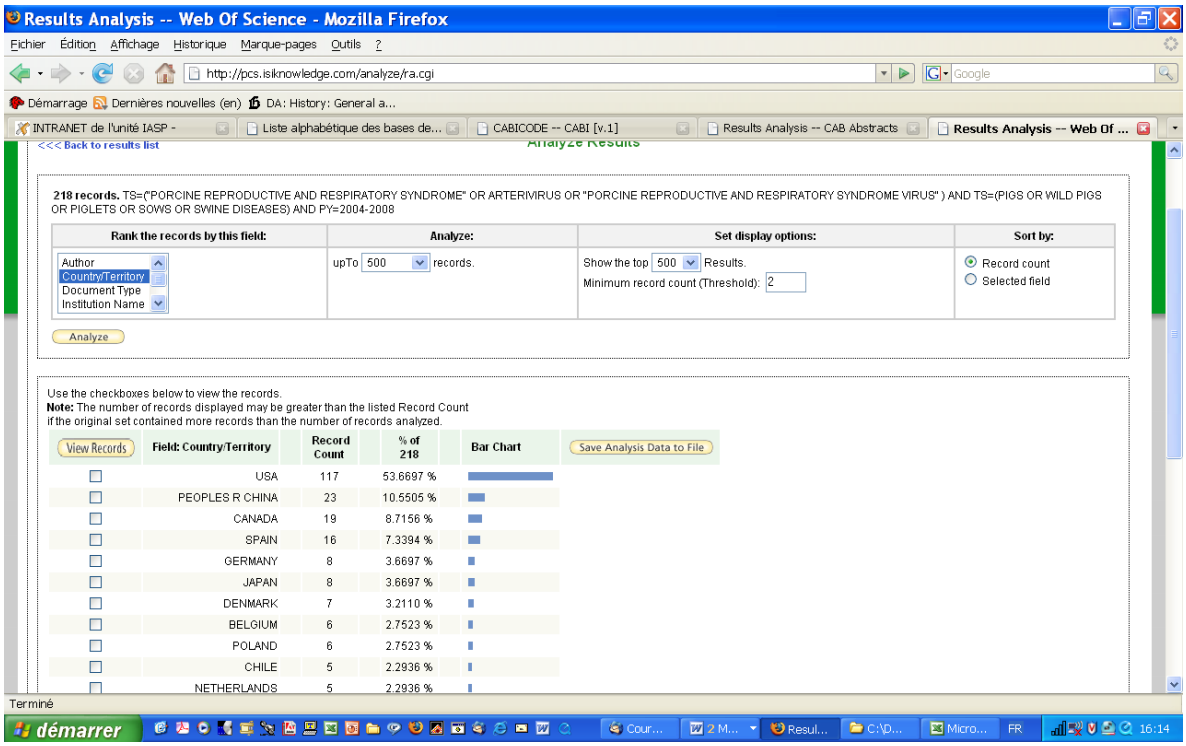
Where "TS" stands for topic and " PY" for years of publication. Note that the disease complex is described by the set of descriptors identified in the previous step, and the production animal category also by a comprehensive set of descriptors picked out from the CAB thesaurus. A WoS advanced boolean search is launched as shown in the screen below :



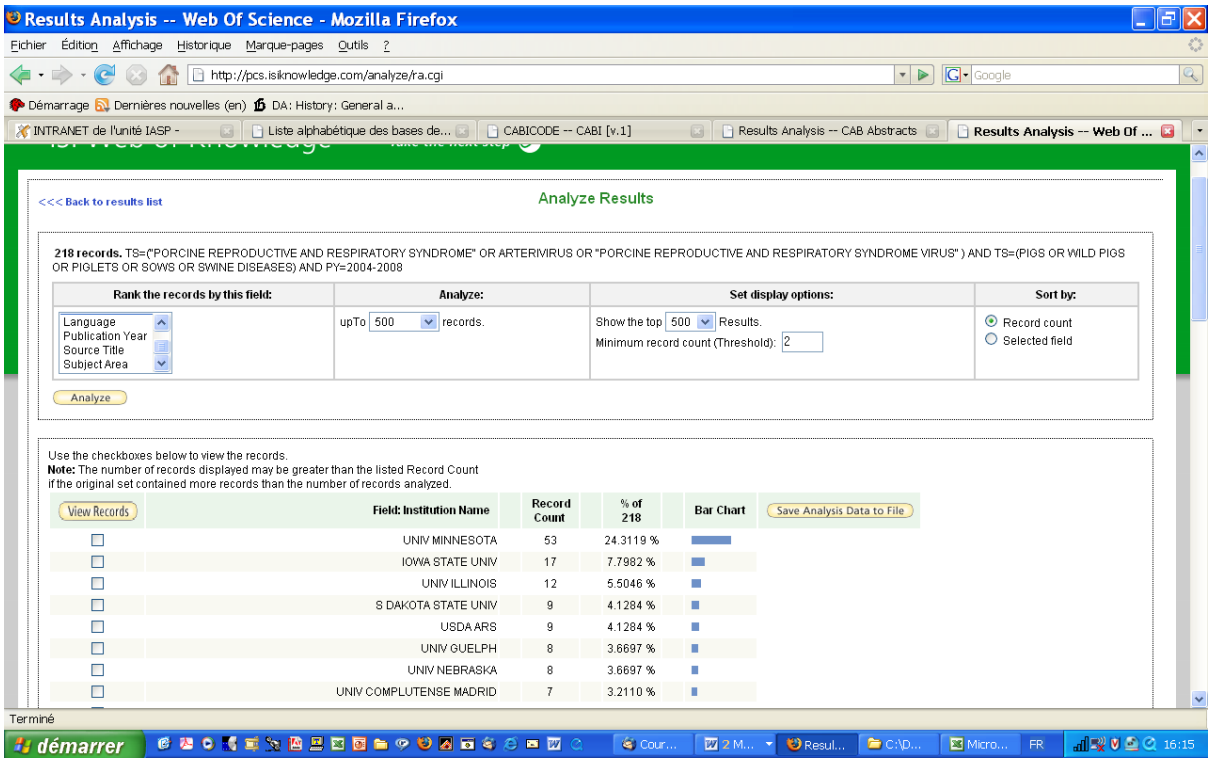
The first results of such a search appears as follows :



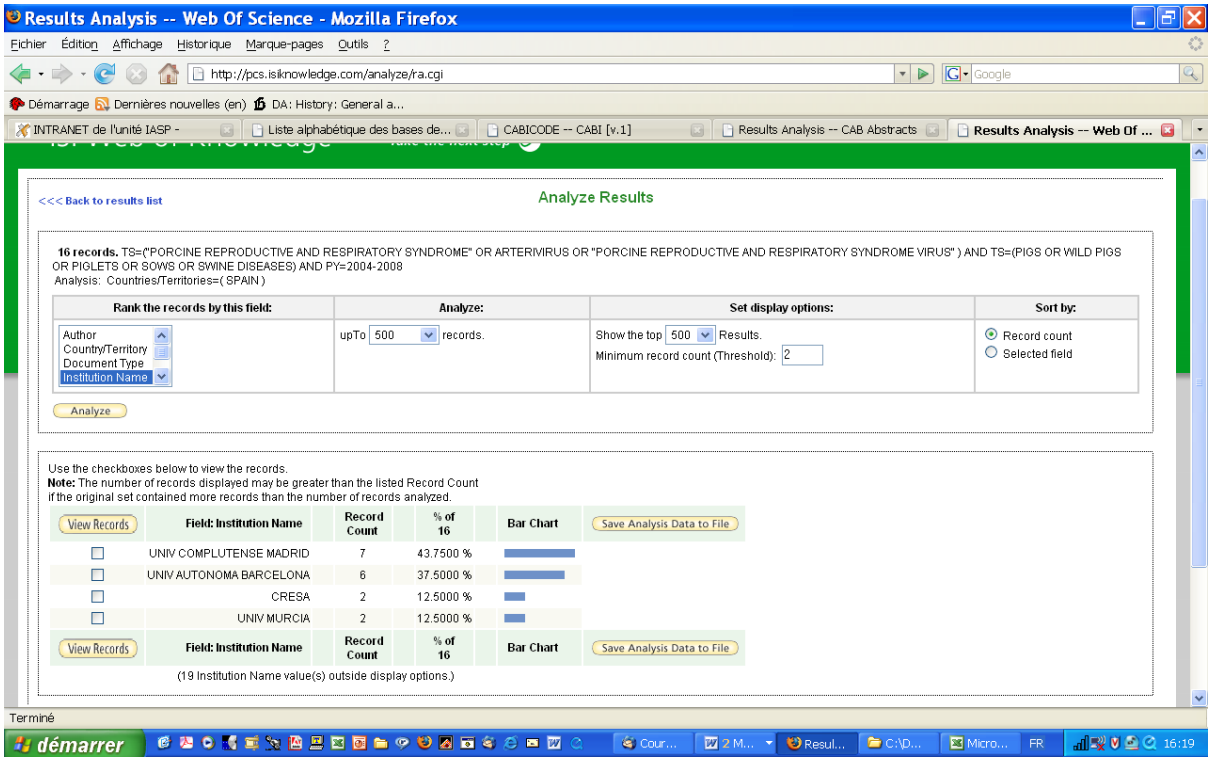
And analysed statistically according to countries :



and to research organizations:



Using the "refine" function one can display the table of institutions working on a specific disease or pathogen in specified countries, and their respective number of publications during the period. As an example, here is for Spain the list of organizations working on the porcine reproductive and respiratory syndrome of pigs :



For EMIDA this procedure has been applied to all countries belonging to the European continent, plus Turkey, as a EC candidate country, and Israël as associate partner of the project: AUSTRIA, BELARUS, "CZECH REPUBLIC", HUNGARY, MOLDOVA, POLAND, SLOVAKIA, UKRAINE, ESTONIA, LATVIA, LITHUANIA, DENMARK, FINLAND, ICELAND, NORWAY, SWEDEN, ALBANIA, "BOSNIA-

HERCEGOVINA", BULGARIA, CROATIA, "REPUBLIC OF MACEDONIA", ROMANIA, SLOVENIA, YUGOSLAVIA, GIBRALTAR, GREECE, ITALY, MALTA, MONACO, PORTUGAL, "SAN MARINO" , SPAIN, VATICAN, ANDORRA, BELGIUM, "IRISH REPUBLIC" , UK, FRANCE, GERMANY, LIECHTENSTEIN, LUXEMBOURG, NETHERLANDS, SWITZERLAND, RUSSIA, TURKEY, ISRAEL

6. Extraction of information on European organisations from the WoS and integration in the EMIDA Excel data base

This information is extracted from the WoS tabulated outputs above, only for those organizations with at least 2 papers during the period surveyed. As the organizations of all the authors of a given source document are identified in the WoS data base, this threshold of 2 papers appear reasonable to select organizations with a significant impact in the topic searched.

In one step, the selected WoS tabulated data are copied into an Excel spreadsheet containing eventually the following fields upon completion of the procedure:

1. **Search query**
2. **Pathogen/ Disease**
3. **Animal**
4. **Source**
5. **Date search**
6. **Number of publications**
7. **Institution Name**
8. **Town**
9. **Country**
10. **Country code (3 Letters)**
11. **Country code (2 Letters)**

7. Standardisation of organisation names

The WoS platform has a standardized procedure for abbreviating institution names of all co-authors of a given publication. The developed name can therefore be easily restored.

However, a same institution can be referred to by two or more different names names in different publications. It was therefore essential to standardise these names so that one institution is referred to by one and single name in the database.

For all the countries that participate in EMIDA, this standardization was entrusted to partner representatives of the project. For the other countries, this task was done by the designer of the database.

8. Specific CAB search for extraction of data on transversal fields

Complementary search on specific transversal fields is possible in the CAB abstracts data base, due to the existence of CABICODES (CC), which indicate the broad subject areas of the publications, and/or of controlled descriptors (DE) to complement CABICODE information.

Search procedures

Considering the opportunities offered by these CAB functionalities, the five following transversal fields were searched, with their associated search query:

Immunology of food animal species

CC=LL650 AND DE=…(animal)……

Genetic resistance to infectious diseases

CC=(LL240 AND (HH600 OR LL821 OR LL822 OR LL650)) AND DE =….(animal)…..

Quantitative epidemiology of infectious diseases

CC=(LL821 OR LL8222 OR LL823 OR LL800 OR YY700 OR EE117 OR LL650) AND DE=("RISK FACTORS" OR "MATHEMATICAL MODELS" OR "SIMULATION MODELS" OR "RISK ANALYSIS" OR "ECONOMIC IMPACT" OR "STOCHASTIC MODELS" OR "ECONOMIC ANALYSIS" OR "COST BENEFIT ANALYSIS" OR "TEMPORAL VARIATION" OR "AGRICULTURAL ECONOMICS") ….

Veterinary vaccines

DE=vaccine* AND CC=((LL821 OR LL822 OR LL823 OR LL650 OR HH600 OR LL882) NOT VV*) AND DE=…(animal) ….

Resistance to antibiotics

CC=(HH410 AND LL821) ….

Resistance to antiparasitic drugs

CC=(HH410 AND (LL822 OR LL823))….

Wildlife parasitic diseases

CC=(YY700 AND (LL822 OR LL823)) ….

Wildlife bacterial diseases

CC=(YY700 AND LL821) ….

Restriction to European countries

To this effect the names of the relevant European countries is introduced in the address field search of the CAB query. Here is the list of the countries:
AUSTRIA, BELARUS, "CZECH REPUBLIC", HUNGARY, MOLDOVA, POLAND, SLOVAKIA, UKRAINE, ESTONIA, LATVIA, LITHUANIA, DENMARK, FINLAND, ICELAND, NORWAY, SWEDEN, ALBANIA, "BOSNIA-HERCEGOVINA", BULGARIA, CROATIA, "REPUBLIC OF MACEDONIA", ROMANIA, SLOVENIA, YUGOSLAVIA, GIBRALTAR, GREECE, ITALY, MALTA, MONACO, PORTUGAL, "SAN MARINO" , SPAIN, VATICAN, ANDORRA, BELGIUM,"IRISH REPUBLIC" , UK, FRANCE, GERMANY, LIECHTENSTEIN, LUXEMBOURG, NETHERLANDS, SWITZERLAND, RUSSIA, TURKEY, ISRAEL

Retrieval of institutions addresses

In contrast to WoS, it is not possible to sort data according to institutions names in CAB. For each specific search, all records were therefore copied in an EndNote file to be sorted by institution name. Besides, as CAB records contain the names and addresses of the first authors of papers only, we selected all institutions and not only those with at least two papers in the period searched (2004-2008).

9. Translation of spreadsheet in web-based searchable database

The output spreadsheet was then turned into a web searchable database after importation using SQL language (Yeoconcept ®). The resulting search interface was being made available on the EMIDA website in April 2009. Simplified selection and sorting of information can be performed starting from the following entries : Topics, Institutions, Countries. Institutions can also be located on maps.

10. Potentials and limitations of the methodology used

The information on recent research output was drawn from the ISI Web of Knowledge ® platform, namely the CAB Abstracts® and the Web of Science®. Some features of our original collection process should be recalled here to point out some potentialities but also some limitations of this methodology.

(i). Thematic coverage of EMIDA research field. The main challenge, before any extraction of data, was to define the field of EMIDA (animal infectious diseases) through a complete set of search queries purported to exhaustively cover this field without major omissions. This challenge was made possible by two inner functionalities of the CAB Abstracts ®, a platform specialized in agronomy and animal health: (i) broad descriptors (the "cabicodes") indexing specific sub-fields of animal health and (ii) statistical analysis of specific controlled descriptors on diseases, pathogenic agents, and animal host names, present in a specific thesaurus.

These particular assets of the CAB ® make it unlikely (but not impossible) that major topics (i.e. topics with significant number of publications in the recent years in Europe) escaped our attention, provided that they be referenced in the source platform. However minor research topics (precisely with less than 2 records referenced in the WoS ® during the 3.5 years period surveyed and at world scale) can obviously be absent. It could be interesting, at the stage of gap analysis, to identify missing topics and analyse if their absence in Europe can easily be accounted for.

(ii). Construction of search queries and disease clustering. Search queries were constructed by grouping together keywords that pertain to the same disease or "disease cluster" in a given animal category. Although performed by an animal health specialist (JDR, the author of this study), this clustering may in some cases be considered excessive, or even unjustified from a mere scientific standpoint. To fully clarify this issue and prevent any misinterpretation, all the query searches used to extract data from ISI Web of Science ® will be accessible in the EMIDA output database. Comments on the query search formulations and on the descriptors grouping scheme will thus be possible on the part of database users to improve any next version of EMIDA output database.

(**iii). Transversal research fields.** Survey of these transversal fields was made reliable thanks to the existence, in CAB Abstracts ®, of broad descriptors covering a large part of them. The combination of the 9 transversal fields presented in Annex VII covers about 60% of the CAB records that were otherwise selected by the combination of the three "cabicodes" defining the field of infectious and parasitic diseases of animals (LL821, LL822 and LL823). In spite of what can be considered as a satisfactory coverage of applied and finalized research topics, major generic topics dealing with basic research, such as pathogenesis and the molecular dissection of pathogenic agents, are missing. The analysis of research data on these topics is essential, in particular with a view to identify new research fronts, and improved approaches should be found in the next phase of EMIDA WP2 to study more efficiently output in basic research in animal infectious diseases.

(iv). Period surveyed. The period of survey starts from 2004 so as to take into account recent research output up to the date of collection. For comparison, a term of 4 years corresponds roughly to the minimum time of completion of a large research programme. The majority of information originated from the ISI Web of Knowledge ® was collected in July 2008, thus covering a 4.5 years term. The specific analysis on publication sources or on co-authorship (section K and M3, respectively) theoretically include data of the full five years 2004 to 2008, having been collected during the fist trimester of 2009. In any case, the dates of extraction are indicated as required in the report, and all the comparisons

presented in this study were done from data sets collected contemporaneously. It is also worth mentioning that the information available in ISI platforms is any kind constantly evolving, even for years already completed (here 2008).

(v). Record counts. In its conception, the primary aim of this database is qualitative and not quantitative. Although counted records reliably reflect the volume of the research output for each entry (topic, country, institution), some features of the data collation process should be recalled to prevent any over-interpretation of such quantitative information.

➔ the original "entry" of data extraction process from the WoS ® is not the publication itself but the research institution. For any given entity (country, topic, research institution), the cumulated "number of records" corresponds to the cumulated number of lines to which this entity is associated in the EMIDA database, and not strictly to the number of publications during the period surveyed. This is due to two main reasons:

- any co-authored publication provides as many lines (or records) as the number of author institutions of the publication;
- many publications are related to both a "disease topic" and one (or even several) "transversal topics", thus generating as many lines in the database.

➔ extraction in WoS has been restricted to institutions with at least two publications during the 3.5 year period of survey. However, it appears that an institution name can have several different wordings in the WoS ®, resulting in the splitting of records under these several names. Consequently, and in spite of careful attention, a few number of eligible institutions (i.e. with two publications or more) may have been overlooked.

----------END OF DOCUMENT----------