The impact of geographic location on the development of a specialty field: A case study of Sloan Digital Sky Survey in Astronomy*

Fidelia Ibekwe-SanJuan

Department of Information & Communication, Jean Moulin University,
4, cours Albert Thomas, 69008 Lyon, France, <ibekwe@univ-lyon3.fr>

Fidelia Ibekwe-SanJuan is currently a professor at the Information-Communication Department of the Jean Moulin University in Lyon, France. Her research revolves around the use of linguistic, terminology methods and the application of language technologies in text mining and information retrieval. Of particular interest to her is how new technology and tools can assist the creation of advanced forms of knowledge organization (KO) systems (topic maps, ontologies). She is also interested in epistemological questions in KO and how research in KO intersects with other fields necessarily engaged in designing and using various forms of KO tools such as Artificial Intelligence, LIS, Natural Language Processing (NLP) and Information Retrieval (IR).

Ibekwe-SanJuan, Fidelia. **The impact of geographic location on the development of a specialty field: A case study of Sloan Digital Sky Survey in Astronomy.** *Knowledge Organization, 35(4)*, 239-250. 19 references.

Abstract: We analyze the scientific discourse of researchers in a specialty field in Astronomy by examining the influence that geographic location may have on the development of this field. Using as a case study the Sloan Digital Sky Survey (SDSS) project, we analyzed texts from bibliographic records along three geographic axes: US-only publications, non-US publications and international collaboration. Each geographic region reflected authors affiliated to research institutions in that region. International collaboration refers to papers published by both US-based and non-US based institutions. Through clustering of domain terms used in titles and abstracts fields of the bibliographic records, we were able to automatically identify the topology of topics peculiar to each geographic region and identify the research topics common to the three geographic zones. The results showed that US-only and non-US research in SDSS shared more commonalities with international collaboration than with one another, thus indicating that the former two focused on rather distinct topics.

* This is a longer and re-worked version of a paper presented at the 10th ISKO international conference, 5-8 August, Montréal, 2008.

1. Introduction

It is a reasonable assumption to think that geographic location can play a determining role in the complex processes involved in knowledge creation, acquisition and organization. However, this parameter has rarely been the focus of automated methods and systems for knowledge representation. It becomes crucial to integrate this dimension when dealing with knowledge that can affect the performance of services at the individual, community or national level. In this study, we aim to investigate how geographic location influences the constitution of a specialty research field. Using an automatic topic mapping system aimed at assisting users in acquiring knowledge from large datasets, we highlight geographic differences in the original data. We take as a case study publications from the Sloan Digital Sky Survey (SDSS) project in Astronomy. The SDSS project aims to collect high quality data for astronomical research and is mostly funded by US institutions such as the NASA and the National Science Foundation.

The SDSS project is a relatively recent one. Begun in 1991, it only started yielding publications since 1998 following the first data release from telescope observations of the stellar objects in the universe. The SDSS project aims to map a quarter of the sky, thus furnishing astrophysicists with 3D images of more than 100 million celestial objects (such as stars, quasars, and galaxies) and spectra of the million brightest galaxies. SDSS project makes regular data releases so that anyone can access the survey data. The publicly-available datasets include not only the images and spectra, but also a database of measured parameters, such as position, brightness, color. The SDSS project has led to a rich emerging literature and a digital record of queries to the data repository (skyserver.sdss.org). The availability of this data has led to an increasing number of discoveries such as high-redshift quasars and significant breakthroughs in astronomical research such as the detection of cosmic magnification caused by the gravitational effect of dark matter throughout the universe. The SDSS project has made important new discoveries in Astronomy. According to the project website, in 2006 alone, it has enabled discovery of "new dwarf companion galaxies to the Milky Way, confirmed Einstein's prediction of cosmic magnification, observed the largest known structures in the universe; and further unraveled our galaxy's active past, filled with galactic mergers" (see http://www.sdss.org/background/).

Given that the SDSS project is mainly funded and operated in the US, a natural question arises about the impact US-based research institutions may have in shaping the structure of this field. In other words, we seek to determine if prominent research themes undertaken by astronomers based in the U.S. differ significantly from their counterparts in other countries and regions such as Europe and Asia. The research questions to which we try to bring answers are: what scientific discoveries made by the SDSS community worldwide can be distinguished along geographical dimensions? What is the overlap between topics in US-based publications and non-US based ones?

2. Methodology

We address these questions from the perspective of the automatic analysis of scientific literature of publications produced by discourse communities related to the SDSS project. Publications from SDSS researchers worldwide constitute communication acts from the same discourse community as they are bound by the same research object in the sense defined by (Swales 1990) and cited in (Borg 2003): "discourse communities are groups that have goals or purposes, and use communication to achieve these goals." We seek to characterize their terminology by an in-depth analysis along geographical axis. Terminology is particularly relevant to the focus of the current study as it will enable us to carry out a detailed study of focus in the scientific discourse in the three data sets (Fellbaum 1998, Nenadic et al. 2004). The type of results produced by our system are research topic maps and terminology network. These constitute knowledge organization artifacts which can be used by specialists in a given field to perform other knowledge organization tasks. The interdisciplinary nature of knowledge organization (KO) as a field of research has been underlined by several authors (see for instance, the special issue of the current journal dedicated to the foundations of KO, edited by McIlwaine & Mitchell 2008). Several definitions of the field have been offered, some contrasting with the others. One component of KO on which all the authors seem to agree is that, among other things, KO is about designing knowledge organization systems (classification schemes, thesauri, subject headings, lexicons, etc) and applying them to index and to retrieve documents. As pointed out by Hjørland (2002) and cited in López-Huertas (2008), amongst the various methods for accessing domain knowledge, terminology analysis plays a vital role because it can reveal the emergence of new terms correlated with new concepts in a domain. Other methods are bibliometric analysis and the joint application of methods (bibliometric, terminological, indexing, etc.). Our approach offers a combination of the first two—bibliometric and terminological, with a particular emphasis on the terminological level.

As more specialized digital collections become available, there is a need to support more advanced and customized access to information especially for domain specialists. For this category of users, information needs, knowledge acquisition and organization are expressed in terms of more advanced computer-assisted representation of the available knowledge stored in electronic memories. One important technique used for discovering and organizing topics from a collection of texts is clustering (Jardine and Van Rijsbergen 1971, Hearst 1999). Clustering offers a means of structuring domain topics and thus furnishes the end user with some sort of map and taxonomy of major domain concepts (Schneider and Borlund 2004). These enhanced forms of domain

knowledge organization are useful when a global view of the domain structure and dynamics is required.

Although, a number of bibliometric tools exist for co-citation analysis and knowledge domain mapping, they are mostly focused on author or journal co-citation data (Small 1999, White & McCain 1998). Few bibliometric tools have considered mapping the content of scholarly communication and when they do, they usually consider the texts as a bag-of-words and ignore the syntactic structure and relationships between the terms. Thus none of the existing bibliometric tools is adapted to the goal of our analysis here, which was to examine the differences or similarities in research topics by a linguistically-oriented processing of the text fields in the underlying bibliographic records. To fill this gap, we developed TermWatch, a topic mapping tool based on Natural Language Processing (NLP) of texts to extract domain terms, establish semantic relations between them and using these relations, cluster them into domain topics. TermWatch integrates state-of-the-art techniques for automatic text data analysis from terminology & natural language processing (NLP), clustering and mapping techniques. TermWatch has been used in a number of topic mapping and terminology structuring studies (SanJuan & Ibekwe-SanJuan 2006, Ibekwe-SanJuan 2006, Ibekwe-SanJuan 2002). It is particularly adapted to topic analysis at the microscopic level, i.e., at the level of content analysis from a corpus of texts.

Research topics are identified by applying shallow NLP techniques to the title and abstract fields of SDSS-related publications. First multi-word terms are extracted. These are nominal phrases (NPs) which can be simplex like "bread basket" (a head-modifier pair) or complex ones such as "wicker bread basket." The latter can be split into two simple NPs "wicker basket" and "bread basket." Then terminological variations are identified in order to establish a network of domain terminology (see §4.2 for some examples). This terminology network is then clustered in other to produce clusters of domain topics. The maps generated by TermWatch reveal the topology of research topics in each geographic region and allows the users to view how the field is structured.

Next, we perform a comparative analysis of the topic obtained based on the map generated for each geographic region, and quantify their overlap. This enables us to identify commonalities and differences in research topics along geographic regions. Our overall methodology can be represented by figure 1.

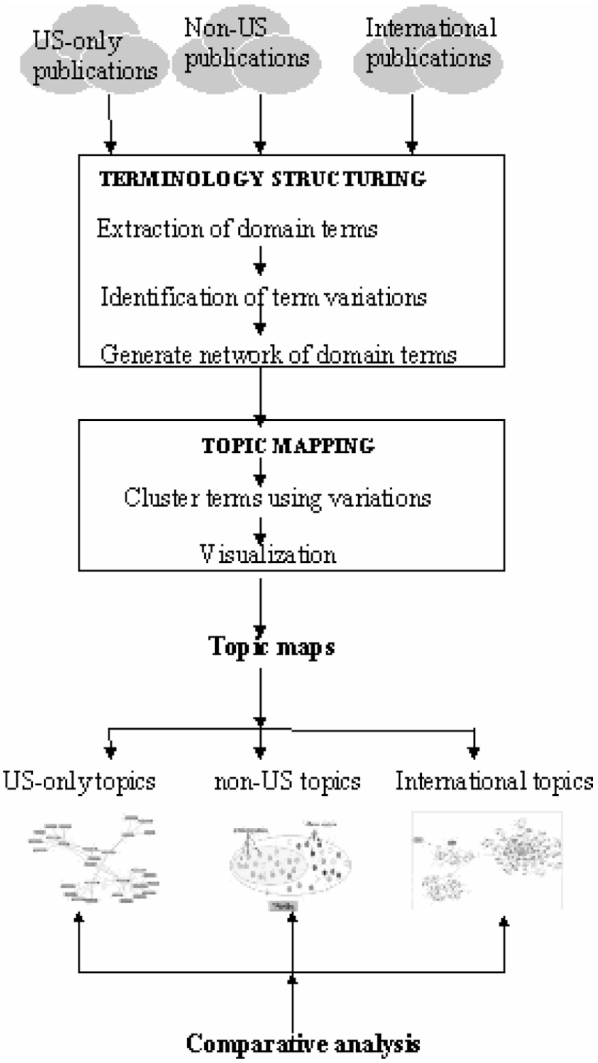

Figure 1. *Flowchart view of the comparative analysis methodology.*

3. Data collection and partitioning

Our data consists of bibliographic records of peer-reviewed journal publications on SDSS between 1998-2007. These records were collected following a search on the *Web of Science* (*WoS*) (http://scientific.thomson.com/products/wos/). A total of 1456 bibliographic records were obtained. The corpus was then split using the affiliation field of the *WoS* records (ISI), i.e. the country in which the research institution is located. Thus, the affiliation field is used to partition the corpus into three subsets: US-only publications, non-US publications and international publications. US-only publications refer to those in which the affiliation field contained only US-based institutions. Non-US publications refer to the opposite case: the authors were affiliated to institutions in different countries except the US. International col-

laboration refers to collaborations between authors from US-affiliated institutions and institutions in the rest of the world. Among the 1456 records, 379 were published by US-based research institutions only, 459 by non-US institutions and 618 were publications between US and non-US institutions (international collaboration). The histogram here below gives a visual image of this distribution.

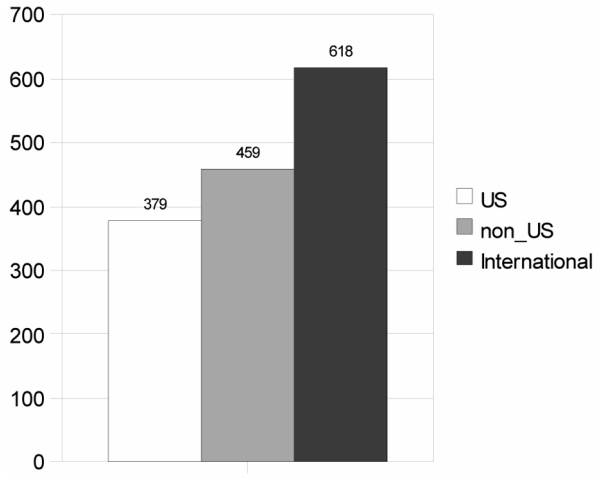

Figure 2. *Histogram of publications by US-only, non-US and International collaboration.*

We can see from these figures that the US-only publications in SDSS journal publications is almost equal to the quantity produced by non-US (the rest of the world). Thus, it is legitimate to seek to determine the impact of the US in shaping the research landscape in SDSS.

4. Domain terminology acquisition and representation

In this section, we briefly outline the processes leading from terminology extraction, terminology structuring to research topic mapping.

4.1 Multi-word Term Extraction

After the corpus has been tagged using TreeTagger (Schmid 1999), contextual rules are used to extract multi-word terms based on morphological and syntactic properties of terms. One such rule is the following:

<mod>* <N>+ of <mod>* <N>+
<prep1> <verb> <mod>* <N>+
then return:
1) <mod>* <N>+ of <mod>* <N>+
2) <mod>* <N>+

where:
<mod> = a determiner (DT) and/or an adjective (JJ)
<N> = a noun tag
<prep1> = all other prepositions excluding "*of*"
* = Kleene's operator (zero or *n* occurrences of an element)
+ = at least one occurrence of an element

This rule favours the extraction of terminological noun phrases in a preposition structure where the preposition is "of." This preposition has been found to play an active role in the formation of multi-word terms. About ten such rules were sufficient to account for nominal composition in English.

4.2 Generating a graph of semantic term variants

We studied linguistic operations which are domain independent and can be used to build taxonomies, thesauri or ontologies in English. Semantic relatedness here is defined as a function of morphological, lexical and syntactic properties shared by some terms. These operations, called terminological variations, stem from two main linguistic operations: lexical inclusion and lexical substitution. By lexical inclusion, we refer to the case where a shorter term is embedded in a longer one through three specific operations: insertions (*severe poisoning* → *severe food poisoning*), modifier or head word expansion (*disaster intervention* → *disaster intervention call*). By lexical substitution, we refer to the case where terms of identical length share a subset of lexical items save one in the same position (*political violence threat* → *political violence campaign*).

Lexical inclusions engender hypernym/hyponym (generic/specific) relations between terms while lexical substitutions indicate a loose kind of semantic association between terms and are by the far the most frequent relation type. Identifying these operations between terms is a way of acquiring semantic relations between them. Lexical substitutions between binary terms give rise to a highly connected graph of term variants which may include some amount of noise (spurious relations). They are filtered using two criteria: we retain only those substitutions that involve terms of length ≥3 if the words in the same grammatical position are found in the same WordNet synset (Fellbaum 1998).

We also acquired explicit synonymy links between multi-word terms using WordNet. To do this, we extended the single word-word relations in WordNet

to multi-word terms by adding these restrictions: two multi-word terms are considered to be in a synonymy relation if two of their words are in the same WordNet synset, occupy the same grammatical role in the terms (both head or modifier words) and are found in the same position. The table below shows some of the synonyms identified in this way. These variations are used in the next stage of processing to form research topics.

Variation type	Term	Variant
Spelling variants	cold-dark-matter model	cold dark matter model
WordNet synonyms	spectroscopic *study*	spectroscopic *survey*
Modifier expansions	hubble deep field	hubble *ultra* deep field
Head Expansions	star formation	star formation *truncation*
Modifier Substitutions	*AGN* luminosity function	*r-band* luminosity function
Head substitutions	recent star formation *activity*	recent star formation *history*

Table 1. *Examples of semantic term variants identified in the SDSS corpus.*

4.3 Term clustering and topic mapping

After term variant identification, terms are clustered based on the variation relations described above. The linguistic significance of each relation can be translated in terms of one of two possible roles: COMP and CLAS. Ideally, COMP relations are variations that induce near-semantic equivalence or synonymy links such as spelling variants, permutations, WordNet synonyms, modifier expansions and insertions. COMP relations are used to form a prior category of tight semantic clusters which serve as a first level of aggregation. The system draws an edge (a link) between two nodes (two terms) if one is a COMP variant of the other. Thus, we first group together terms for which there is a sequence of variations in COMP. Since variations in COMP are supposed to link only closely semantically related terms, resulting connected components are topically coherent, i.e., reflect different properties of the same concept. Components are labeled by the most active term. Prior grouping of term variants into components ensures that semantically close terms which reflect different aspects of the same topic are certain to end up in the same cluster at the end of the process.

CLAS relations are involve a topical shift between two terms, i.e., where the head word is different like head expansion and head substitution. This category of relations is used to cluster the components formed by COMP relations in a hierarchical process using the weight of CLAS relations between each component. TermWatch chooses as cluster label, the term with the highest number of variants. This term can be considered a good representative of the class.

This way of regrouping terms either by shared modifiers or by shared head is known as distributional analysis and was introduced by Harris (1968) and later taken up by studies on automatic thesaurus construction (Grefenstette 1997, Wacholder 1998). We extended the definition of the types of relations identified and added additional constraints such as the position of added words and their number to avoid generating spurious variants. A more formal description of the clustering algorithm can be found in SanJuan & Ibekwe-SanJuan (2006). Table 2 gives an example of a cluster.

Cluster label : Quasar luminosity
Contents: Luminosity function, AGN luminosity function, AGN luminosity, band galaxy luminosity function, cluster luminosity function, composite luminosity function, derived luminosity function, emission-line luminosity function, galaxy luminosity function, local galaxy luminosity function, k-band luminosity function, local x-ray luminosity function, x-ray luminosity function, observed luminosity function, QSO luminosity function, radio luminosity, radio luminosity function, quasar luminosity, quasar luminosity function, r-band luminosity function, schechter luminosity function, cluster LF, line luminosity. Accurate photometry, weighting scheme, strong dependence, flatter slope, composite LF.

Table 2. *Example of a cluster (research topic) generated on the SDSS corpus.*

The majority of the terms grouped into this cluster are semantic variants of "luminosity function" automatically identified by the system. We observe that this generic term has been abbreviated by authors as "LF" and used in the longer terms such as "cluster LF, composite LF." The clustering algorithm was able to capture these semantic variants and group them into the same cluster without manual intervention. Fewer terms in this cluster result from co-occurrence associations (terms on last line), which complements the linguistic dimension for clustering. In most bibliometric systems where items are grouped by co-

occurrence, these semantically related terms would have been dispersed in different clusters.

5. Results

We first analyze the topology of research topics for each geographic region (§5.1). Then, we perform a terminological survey of topics found in each region by a comparative analysis of cluster contents (5.2). This terminological analysis will help us identify overlapping and distinct research topics in the three geographic regions.

5.1 Structure of SDSS research by geographic regions

TermWatch produced maps of research topics for each region: US-only, non-US and International. The system automatically identifies highly connected topics (called central atom) and loosely connected topics (called peripheral atoms). Also the system performs a chronological analysis of these maps by using the publication year of each paper. This is reflected as a color scheme on the nodes (clusters of research topics) to indicate the period in which the terms of that topic appeared. Owing to printing constraints (black and white images only), the color coding system cannot be shown to its full advantage.

5.1.1. Structure of the major topics in the US-only institutions

Three hundred and seventy-nine papers were published by US-only authors. The map below (figure 3) shows the global view of major research topics. This map has a cyclic shape reflecting a highly connected set of topics. Most of the topics were found in the

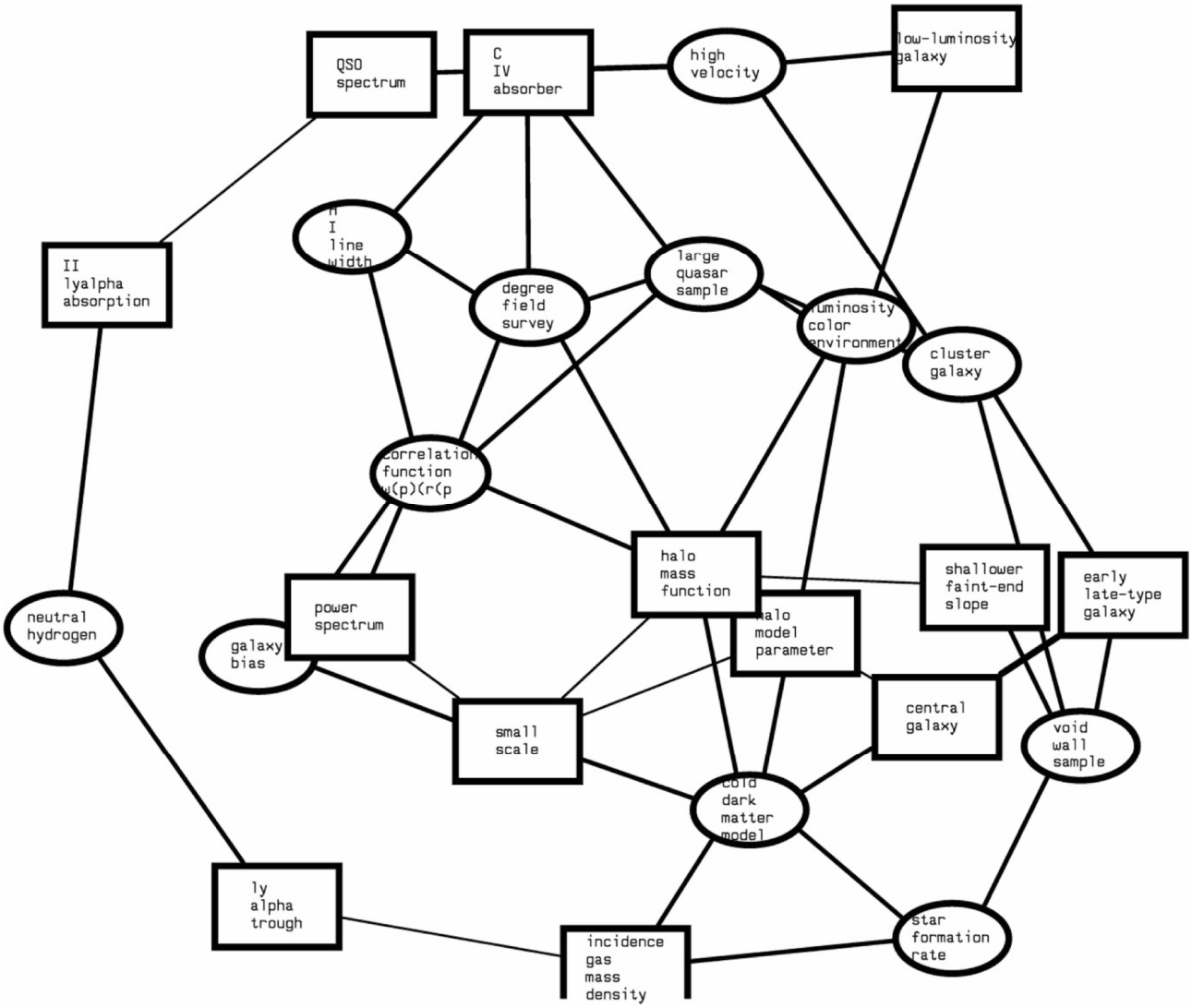

Figure 3. *Global image of major research topics in the US-only publications.*

last period of the corpus (publications made between 2005-2007) and thus were quite recent. Topic labels found in this period are "low luminosity galaxy, cluster galaxy, correlation function, halo mass function, shallower faint-end slope, halo model parameter, cold dark matter model, central galaxy, small scale, void wall sample, star formation rate, incidence gas mass density, neutral nitrogen, ly alpha trough". The most central cluster labeled "halo mass function" is focused on galaxy clustering and formation models basing on the measurement of their halo mass and luminosity functions.

Surrounding clusters deal with measurements and models of galaxies drawn from the SDSS data releases in order to predict galaxy clustering and galaxy evolution. The cluster "central galaxy" refers to the study of the relation between "central galaxy luminosity" and "halo mass," and to the study of the relationship between galaxy luminosity, color, and environment in a cosmological simulation of galaxy formation. Labels found in the mid period of the corpus (2003-2005) are "large quasar sample, luminosity color environment, power spectrum, galaxy bias, II ly alpha absorption, early late-type galaxy." These clusters deal with the detection of quasars, the correlation function of high redshift objects such as quasars, the study of the relation between galaxy luminosity, color and environment. Three clusters labeled "QSO spectrum, C IV absorber, high velocity" refer to research topics that appeared between 1999-2001. The cluster "degree field survey" denotes a topic whose terms peaked in the period between 1996-1998. On the whole, the major research topics in SDSS in the US seem to have a highly inter-connected structure.

5.1.2. Structure of research topics in non-US publications

There are 459 publications in the non-US dataset. Figure 4 shows the global image of topics found in this data set. The topology of the map shows that there is no one central atom as in the US-only research. Research outside the US seem to be organized around five major research topics with its one topic acting as core and connecting the other related topics. This topology may be explained by the fact that non-US publications concern the rest of the world, thus it is more expected that different research directions will be explored in parallel by different research teams in different geographic regions outside the US. Hence a concentration around a unique center is less expected. The map of the major research topics shows an elongated form which cannot legibly be captured in an image view. To obtain a global image view, we had to reduce its dimensions but at the expense of legibility of cluster labels. For ease of analysis, we have labeled the five centers which connect other groups of clusters on the map. These five major clusters are "black hole, seyfert galaxy, star formation rate, supernova type ia, nearby cluster." We explored the contents of these five clusters as well as neighboring clusters. Some of the clusters in the "black hole" group are "black hole mass, broad absorption line, emission line region." "Seyfert galaxy" is linked to the "black hole" group by a cluster labeled "emission line region."

The star formation rate group articulates research around the process of star formation as evidenced by neighboring clusters labeled "stellar mass metallicity, total stellar mass, star formation." This group of clusters is linked to the "nearby cluster" group by the cluster labeled "star formation," showing the proximity of the two groups of topics. Here the term cluster refers to clusters of galaxies. The "supernova type ia" group is linked to a star-shaped group of clusters some of which are labeled "cosmic microwave background shift parameter, dark matter particle, dark matter particle mass, lambda CDM universe."

5.1.3. Structure of research topics in international collaboration

These are publications co-authored simultaneously by US and non-US institutions. 618 records were concerned. Like the non-US research, international collaboration in SDSS is not organized around a unique center. Several groups of research topics are connected through chains of intermediary topics. For the same reasons already evoked, exporting a global image view is at the expense of legibility of cluster labels. We have circled and labeled the cluster at the center of the different groups for legibility reasons: "cosmological parameter, galaxy-galaxy weak lensing, faint end slope, sagittarius dwarf, stellar mass, fifth lensed image, complete gunn-peterson trough."

5.2. Comparative analysis of topics across the three geographic regions

We now carry out a more detailed exploration of similarities and differences among research topics in SDSS in the three geographical regions based on the cluster contents. This comparison is carried out at two levels:

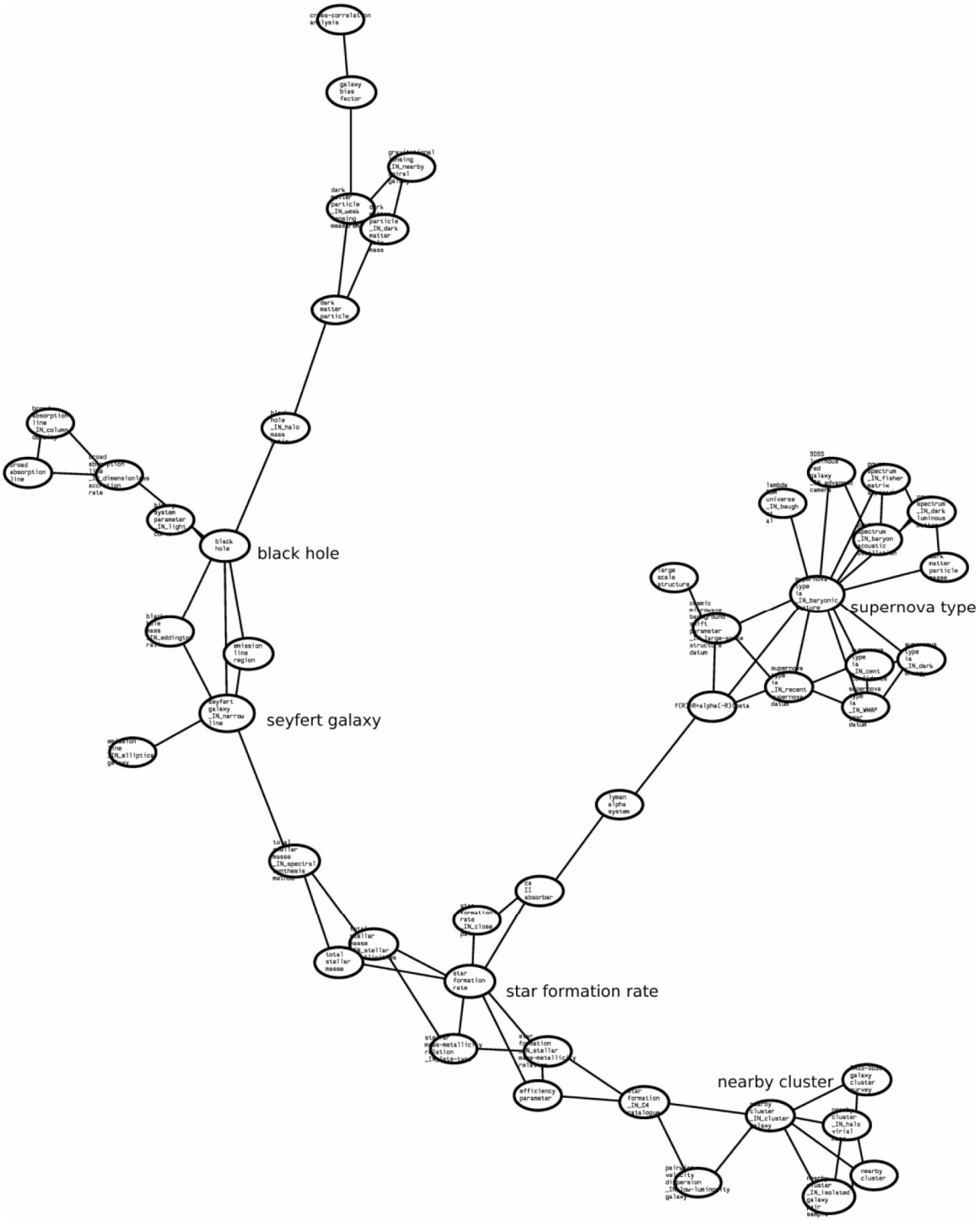

Figure 4. *Map of topics from non-US publications on SDSS.*

– cluster label which is the representative of each research topic (a kind of descriptor)
– cluster content comparison in the three geographic regions.

The idea is to determine if there are research topics that characterize each geographic region and if there are some that are shared by pairs of geographic regions or by all three regions.

5.2.1 Similarities in research topics labels

TermWatch automatically labels its clusters with the most active term in terms of terminological varia-

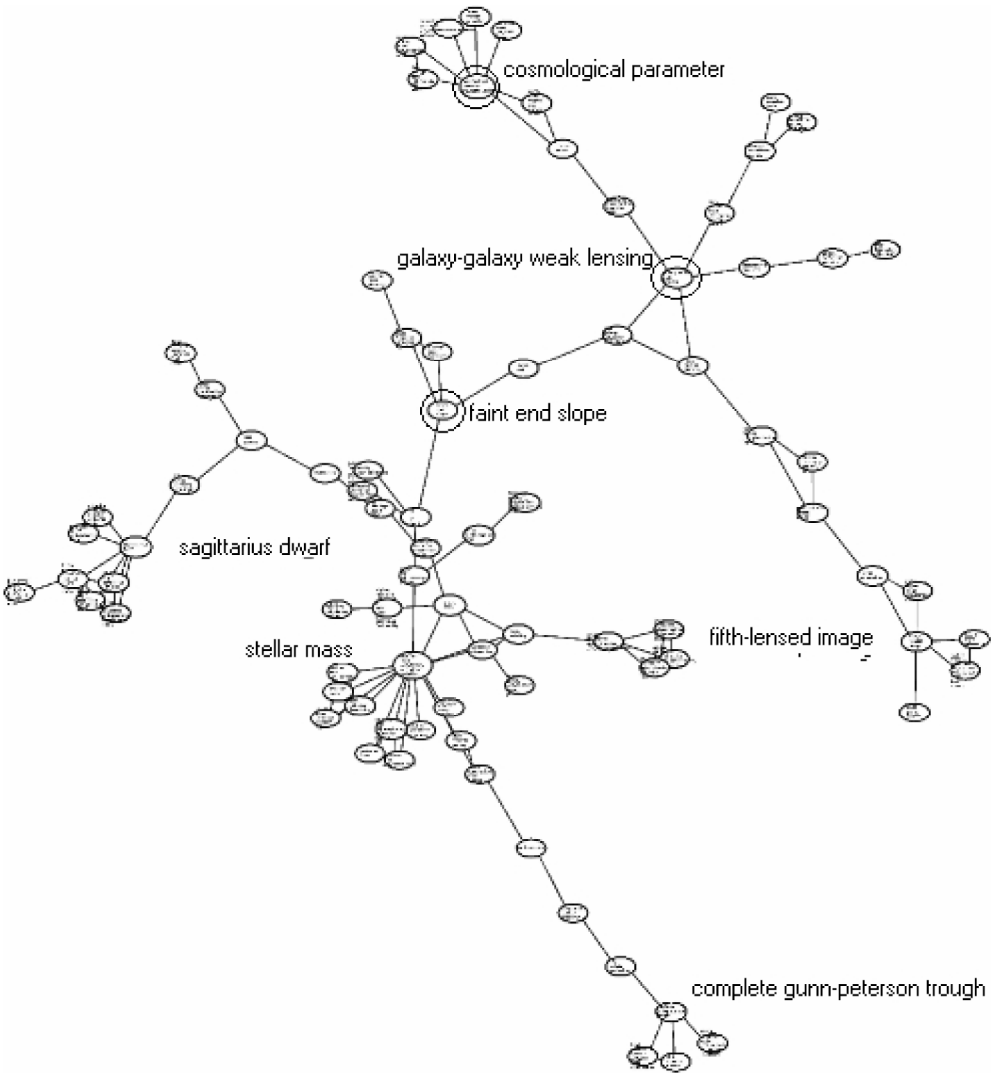

Figure 5. Map of topics from publications in international collaboration.

tions (the term with the highest number of variants). This term can be considered a good representative of the topic. TermWatch generated 163 clusters in the non-US publications, 119 clusters for the US-only and 240 clusters for international collaboration. Table 3 shows the overlap in cluster labels across the three data sets, then for each pairwise set.

The overlap in cluster labels is quite low, thus pointing to significant differences in SDSS research across different geographic regions. It appears from the above figures that both US-only and non-US research share more common points with international collaboration than with one another. Table 4 gives the list of the common labels found. The labels in the first row are common to all three geographic zones and are thus not repeated in their respective rows.

	Non_US	*US_only*	*Inter*
Total nb_clusters	163	119	240

	Total clusters	*Overlap (%)*
US, NonUS, Inter	552	6 (1 %)
US vs Non_US	282	10 (4 %)
US vs Inter	359	22 (6 %)
Non_US vs Inter	403	29 (7 %)

Table 3. *Overlap in cluster labels by geographic and cultural zones.*

From table 4, it appears that the topics of "star formation, emission line, surface brightness, rest frame, large scale structure" are shared by all SDSS researchers regardless of geographic location.

	Topic labels
US, Non-US, Inter (6)	star formation rate, emission line, surface brightness, black hole, rest frame, large scale structure
US, non_US (10)	SDSS spectroscopic datum, power spectrum, cold dark matter model, sloan digital sky survey spectrum
Non-US, Inter (29)	composite quasar spectrum, good agreement, radio-loud, RR lyrae, M circle, scalar spectral index, high-resolution, high redshift quasar, high redshift, power law, cluster mass function, accretion rate, light curve, cosmic microwave background, dark matter halo, BAL quasar, elliptical galaxy, column density, ZZ ceti instability strip, mass density, cold dark matter model
US, Inter (22)	sloan digital sky survey early datum release, low-mass, mock catalog, early datum, radio-quiet, galaxy evolution explorer, dark energy model, early structure formation, mean neutral fraction, lambda CDM model, principal component analysis, line-of-sight velocity dispersion, cold dark matter model, micron all sky survey 2MASS, equivalent width

Table 4. *Common topic labels shared across different geographic regions.*

5.2.2 Similarities in topics contents

Comparison of the clusters contents obtained for each data set gives a measure of their overlap across the three geographic regions. This is a step further because we do not just look at the labels but we also evaluate the proportion of common terms within clusters. The following table gives the details of this comparison.

	Non_US	*US_only*	*International*
Total nb_terms	442	342	683

	Total terms	*Overlap (%)*
US, NonUS, Inter	1467	72 (5 %)
US vs Non_US	784	86 (11 %)
US vs Inter	1025	137 (13 %)
Non_US vs Inter	1125	153 (14 %)

Table 5. *Topic content overlap across geographic and cultural zones.*

The proportion of overlap in topics contents echoes the ones found among topic labels. Thus, similarities are consistent whether we look at the topic labels alone or into their contents. This consistency is remarkable considering that the terms were extracted automatically from the text fields of the titles and abstracts and were not humanly attributed keywords. This term extraction procedure was able to automatically identify the subset of invariant terminology in the SDSS publications across distinct geographic regions. The system was also able to automatically isolate the set of shared knowledge among SDSS researchers worldwide without resorting to a human perusal of the publications which would have been too time consuming. The overlap observed in the three data sets, although small, indicates a certain stability in the terminology employed by SDSS researchers worldwide. Table 6 gives examples of some the common terms.

Terms common across the three geographic regions	
US, Non-US, Inter	black hole, black hole mass, brightest cluster galaxy, cluster mass function, cold dark matter model, cosmic microwave background, dark energy model, dark matter halo, dwarf galaxy, early-type galaxy, galaxy luminosity function, lambda CDM model, micron all sky survey, photometric redshift, quasar luminosity function, ROSAT all-sky survey, specific star formation rate, stellar velocity dispersion, supermassive black hole, wilkinson microwave anisotropy probe
US *vs* Inter	halo occupation distribution, ly alpha system, satellite galaxy, column density, dark energy model, dwarf galaxy, early data release, galaxy evolution explorer, hubble space telescope, local galaxy density, low-mass galaxy, micron all sky survey 2MASS, optical spectrum, principal component analysis, radio-quiet quasar, sagittarius dwarf galaxy, specific star formation rate, spitzer space telescope, large scale structure, velocity dispersion
Non-US *vs* Inter	accretion rate, BAL quasar, candidate RR lyrae, column density, concentration index, dark energy equation, density profile, emission line, gravitational lensing, matter power spectrum, RR lyrae, late-type galaxy, massive galaxy, photometric redshift accuracy, old stellar population, young stellar population, SDSS data release, stellar population, radio-loud quasar, weak gravitational lensing, ZZ ceti instability strip
US *vs* Non-US	axis ratio, brightest cluster galaxy, central black hole, cluster mass, SDSS, correlation function, cluster mass function, early late-type field galaxy, spectroscopic datum, high redshift, micron all sky survey, primordial power spectrum, quasar luminosity, galaxy luminosity function, large scale structure, quasar luminosity function, rest frame, photometric redshift, sloan digital sky survey spectrum, tidal stream, velocity dispersion

Table 6. *Examples of common terms in topics across geographic regions*

5.2.3 Differences in topics by geographic regions

We have so far portrayed similarities both in topic labels and contents. Here we give some examples of differences, i.e., of topics characterizing specific geographic regions and are not found in the other two regions. This enables us to better visualize the differences in research topics in the three data sets.

US	High velocity, high-redshift source, white dwarf-red subdwarf system, incidence gas mass density, large quasar sample, quasi-stellar object, proper-motion measurement, neutral hydrogen fraction, hubble space telescope advanced camera, low redshift universe
non-US	dark matter halo mass, seyfert galaxy, artificial neural network, balmer absorption, high redshift object, three-year wilkinson microwave anisotropy probe, independent component analysis, gaussian initial condition, large-scale structure formation, two-micron all-sky survey 2MASS
International	galaxy-galaxy weak lensing, galactic plane, SDSS optical spectrum, cluster mass profile, RASS SDSS datum, gravitational lensing, automated selection algorithm, SUUMa-type dwarf nova, cosmological parameter, photometric error

Table 7. *Some topics labels specific to publications in each geographic region.*

Although the exact labels in each data is different, we observe that some are semantic variants of terms in the common set. For instance, "dark matter halo" which is a topic label common to non-US and International clusters (table 4), is a more generic term variant of "dark matter halo mass" found as a label specific to non-US topics. Although we did not find a high overlap of exact terms in the three data sets, the proportion of overlapping concepts might be a much higher if we were to extend this comparison to semantically-related terms.

6. Conclusion

We have mapped out the structure of the SDSS research field based on publication records split along three geographic regions: US, non-US and International collaboration. The specific goal of our study was to automatically identify topics that characterized the three geographic regions and highlight their similarities and differences. We calculated overlap of exact terms in research topics and found this to be consistently low whether we were looking at topic labels or contents. The low level of overlap would suggest that geographic location does indeed have an influence in the choice of research topics in a given field. The three geographic zones we examined have many more unique terms characterizing their research topics than common terms. This is more evident for US *vs* non-US research. More expectedly, US-only and non-US topics had slightly higher level of overlap with topics from international collaboration. This tends to indicate that research in SDSS within and outside the US are brought together by international collaboration.

However, we already observed that our comparison in terminology overlap was a strict one because we were only looking at the overlap of exact terms and not their semantic variants. The overlap may be much higher if we relaxed the criteria to include semantically-related terms, i.e; synonyms, hyponyms/hypernyms, associated terms. In such a case, we may observe a more connected structure for the three geographic regions, thus less distinct research. This is a matter for future investigation because it will need a careful selection of particular variations that will preserve the semantic class of a term.

Another significant observation in this study is that the topology of US-only research in SDSS is cyclic while the maps obtained for the rest of the world and for international collaboration showed several distinct subgroups, as if researchers were exploring different avenues in parallel. Let us bear in mind that the maps obtained were the results of wholly automated processes not requiring any human intervention.

Alongside the role of geography on the development of this specialty field, the results produced by TermWatch offer a means of organizing domain concepts in this field according to a user defined axis. In this instance, the system offers maps of topics and a structuring of domain vocabulary. These maps constitute knowledge organization artifacts for researchers in the field. They offer a means of structuring domain terms into classes of related concepts that depict research topics in the field. They can also serve as a starting point to build a specialized taxonomy or thesaurus for a field. For young researchers embarking on research in the field, these maps offer a global view of current trends in the field. The results obtained here are encouraging for identifying the impact and the uniqueness of each geographic region in shaping the SDSS field.

References

Borg, Erik. 2003. Discourse community. *ELT journal* 57: 396-98.

Fellbaum, Christiane. 1998. *WordNet, An Electronic Lexical Database*. MIT Press.

Grefenstette, Gregory. 1997. SQLET:Short Query Linguistic Expansion Techniques, Palliating One-Word Queries by Providing Intermediate Structure to Text, Proceedings of "Recherche d'Information assistée par ordinateur" (RIAO), 500–509.

Harris, Zellig S. 1968. *Mathematical structures of language*. New York: Wiley.

Hearst Marti. 1999. The use of categories and clusters for organizing retrieval results. In Strzalkowski, Tomek, ed. *Natural language information retrieval*. Kluwer Academic Press, pp. 333-74.

Hjørland, Birger. 2002. Domain analysis in information science. eleven approaches–traditional as well as innovative. *Journal of documentation* 58:422-62.

Ibekwe-SanJuan, Fidelia. 2006. Clustering semantic relations for constructing and maintaining knowledge organization tools. *Journal of documentation* 62: 229-50.

Ibekwe-SanJuan, Fidelia and SanJuan Eric. 2002. From term variants to research topics. *Knowledge organization* 29: 181-97.

Jardine, N and Van Rijsbergen, C.J. 1971. The use of hierarchic clustering in information retrieval. *Information storage and retrieval* 7: 217-40.

López-Huertas Maria. 2008. Some current research questions in the field of knowledge organization. *Knowledge organization* 35: 113-36.

Nenadic, Goran, Spasic, Irena and Ananiadou, Sophia. 2004. Mining term similarities from corpora. *Terminology* 10: 55-81.

SanJuan, Eric and Ibekwe-SanJuan, Fidelia. 2006. Textmining without document context, *Information processing & management* 42: 1532-52.

Schneider, Jesper W. and Borlund Pia. 2004. Introduction to bibliometrics for construction and maintenance of thesauri. *Journal of documentation* 60: 524-49.

Small, Henry. 1999. Visualizing science by citation mapping. *Journal of the American Society for Information Science* 50: 799-813.

Swales, John. 1990. *Genre analysis: English in academic and research settings*. Cambridge University Press.

Salager,-Meyer Françoise. 1990. Discoursal movements in medical English abstracts and their linguistic exponents: a genre analysis study. *Interface* 4n2:107-24.

Wacholder, Nina. 1998. Simplex NPs sorted by head: a method for identifying significant topics within a document, Workshop on the Computational Treatment of Nominals, in the Joint 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL'98), Montreal, Quebec, Canada, August, pp. 70-79.

White, Howard D. and McCain, Katherine W. 1998. Visualizing a discipline: an author cocitation analysis of information science, 1972–1995. *Journal of the American Society for Information Science* 49: 327–55.

Zitt, Michel and Bassecoulard, Elise. 1994. Development of a method for detection and trend analysis of research fronts built by lexical or cocitation analysis. *Scientometrics* 30: 333–51.