GREY LITERATURE CITATIONS IN THE AGE OF DIGITAL REPOSITORIES AND OPEN ACCESS

Silvia Giannini, Stefania Biagioni, CNR-ISTI, Pisa Italy
Sara Goggi, Gabriella Pardelli, CNR-ILC, Pisa Italy

Summary

This study investigates the “world” of scientific citations in the areas of Computational Linguistics, Computer Science and Engineering.

- Scenario
- Objective
- Material and Method
- Analysis of Data & Results
- Conclusions

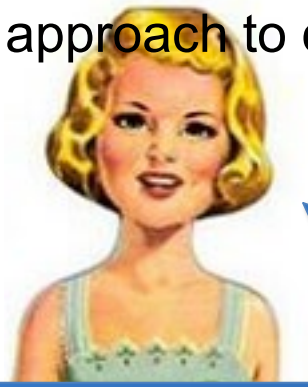
Scenario

- About ten years ago we studied the impact of grey literature (GL) on conventional literature by observing the impact of grey citations in two different scientific fields, «after the growth in the use of the WWW» …
- Over the last decade the international scientific community and its players have undergone (and still undergo) essential changes with respect to the representation and dissemination of knowledge.
- The formalization of the Open Access (OA) model made many academic and research institutions adhere to the OA initiative by issuing policies which compel to deposit/file their research products in OA repositories …
- In our current digital era, bibliographical citations have gained a strategic role within the mechanisms of scientific communication, especially due to the implementation of the citation indexing services.

Objectives

- In this scenario, it seemed interesting to investigate once again the "world" of scientific citations for proving if - and eventually to which extent – this "revolution" in the communication of knowledge might actually reflect on the GL approach to citations…

Two questions…



Has Grey Literature - more easily identified and accessible - a greater visibility *today*?

similarly

Is it possible to assume a greater impact of GL citations on the overall total of citations?

Material and method

This work measures grey citations in the years 2012, 2013 and 2014 (journals) and 2012 and 2014 (proceedings) and then describes the features of GL documents cited in different areas of knowledge: Computational Linguistics, Computer Science and Engineering.

Journals Titles	IF* 2012	Rank* 2012	IF* 2013	Rank* 2013	IF 2014	Rank* 2014
ACM Transactions on Information Systems	1.070	59/132	1.300	53/135	1.021	70/139
EURASIP Journal on Advances in Signal Processing	0.807	155/243	0.808	164/248	0.777	170/249
Computational Linguistics	0.940	72/115	1.468	49/121	1.226	72/123
Language Resources and Evaluation	0.659	79/100	0.518	94/102	0.619	89/102

Sampled Journals

A sample of journals indexed by the Science Citation Index (SCI) and ISI-Journal Citation Report (JCR); and also by Scopus Citation DB (Elsevier)

Proceedings Titles	Years
DL - ACM/IEEE-International Conference on Digital Libraries	2012 and 2014
EACL - Conference of the European Chapter of the Association for Computational Linguistics	2012 and 2014

Sampled Proceedings

A sample of two conference proceedings belonging to different scientific communities.

Material and method

- Information is extracted directly from primary sources, that is, the bibliographical references of the articles in the selected journals and proceedings.

- The obtained corpus contains 40.511 bibliographical references on 1.270 articles
  - including: “editorial”, “obituary”, “squibs”, “book review” “report”, “brief report”, “project note”, “editors’ notes”, “introduction” etc.

- The corpus was built by grouping the gathered data in informative classes:
  - year, issue number, bibliographical reference, kind of document - Grey (G) or Published (P) - document type, standardized document type.

Year	Issue	Reference	G/P	Doc. Type	SDType
2012	38(2)_2	Horn, Laurence R. 1972. On the Semantic Properties of Logical Operators in English. Ph.D. thesis, UCLA. Distributed by the Indiana University Linguistics Club, 1976.	G	PHD	Thesis

Example extracted from Computational Linguistics (2012)

Material and method

For each source, we counted

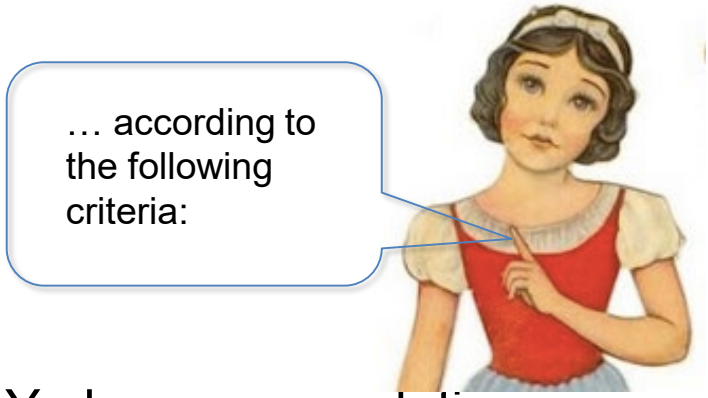
a) the number of articles provided with references;
b) the number of references in each article;
c) the number of GL references in each article;
d) the number of GL references with a URL;

For each GL reference, we examined:

- the document type;
- the year of publication;
- the eventual URL (linked references);

according to the following criteria …

Material and method



1. Definition of GL starting from the York recommendations (1978) and the later integrations to its definition;

2. Classification of documentation produced by *no-profit* Associations, Institutions and Publishers as Grey Literature (e.g. ACL Anthology, ISCA archive , OA journals...);

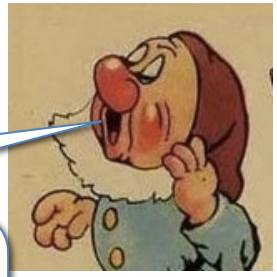
3. Use of specialized indexes, catalogues and Google search to clarify incomplete or unclear citations;

4. Categorization of GL documents typology as follows …

… Criteria

- ***ARTICLE*** includes: journals, newspapers, newsletters and magazines articles;
- ***BLOG/FORUM***;
- ***BOOK/BOOK CHAPTER*;**
- ***CONFERENCE PAPER*** includes: papers presented at conferences, seminars, workshops, meeting;
- ***CORPORA*** includes: downloadable linguistic resources;
- ***COURSE MATERIAL*** includes: tutorials and teaching material;
- ***DATABASE*;**
- ***DATASET*;**
- ***DELIVERABLE*;**
- ***GUIDELINES* and *NORMATIVE DOCUMENT*** includes: standard, guidelines, protocols;
- ***PATENT*;**
- ***PREPRINT/POSTPRINT*** includes: documents “submitted-to”, “to-be-published”, “in press”, “forthcoming”; “accepted”; “to appear”;
- ***POSTER/PRESENTATION*** includes: demo, poster and presentation;
- ***REPORT*** includes: working notes, technical reports, white papers, working papers, research reports, project reports, discussion papers, occasional papers;
- ***SOFTWARE*** includes: only downloadable software;
- ***TECHNICAL DOCUMENTATION*** includes: user guides, manuals, technical specifications and technical documentation of computer programs and for statistical surveys;
- ***TERTIARY DOCUMENT*** include :dictionaries, catalogues and encyclopedia entries;
- ***THESIS*** includes: PhD thesis, dissertations, master thesis;
- ***UNDEFINED*** includes: all documents that could not be identified;
- ***WEBSITE*** includes: simple URLs’ or home pages.

… Criteria



- The different impact of GL on the different areas of knowledge has been analyzed using the following indicators:

1. the frequency of GL citing (i.e. the proportion of GL references out of all the references examined);

2. the frequency of GL use (i.e. the proportion of articles with GL citations, out of all articles examined);

3. the intensity of GL use (i.e. the frequency of GL citing divided by the frequency of GL use).

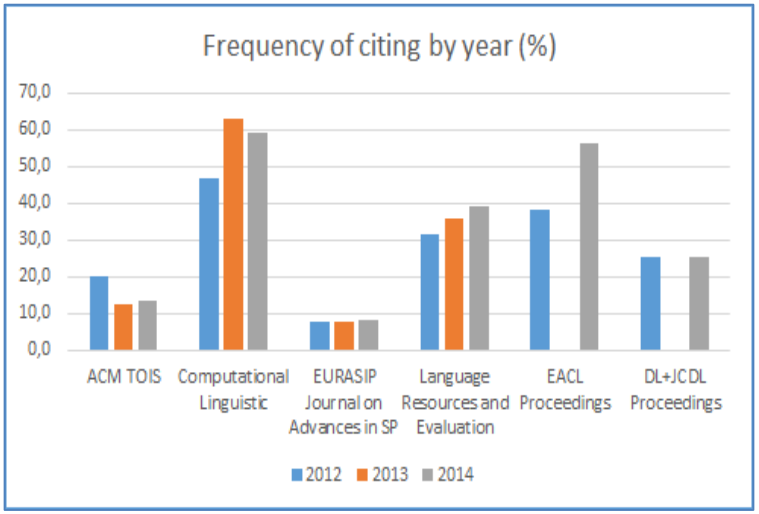
Analysis of data and results

Frequency of GL citing (24% out of 40.511 references)

Frequency of GL citing									
	2012			2013			2014		
Title	Number of references	Number of GL references	Frequency of GL citing (%)	Number of references	Number of GL references	Frequency of GL citing (%)	Number of references	Number of GL references	Frequency of GL citing (%)
ACM TOIS	1.413	285	20,2	1.097	135	12,3	1.096	150	13,7
Computational Linguistic	1.575	739	46,9	2.008	1.263	62,9	1.958	1.158	59,1
EURASIP Journal on Advances in SP	7.876	616	7,8	5.805	459	7,9	5.651	455	8,1
Language Resources and Evaluation	1.220	384	31,5	2.052	740	36,1	1.267	495	39,1
EACL Proceedings	2.307	884	38,3	/	/	/	2.368	1.332	56,3
DL+JCDL Proceedings	1.304	329	25,2	/	/	/	1.514	384	25,4
Total	15.695	3.237		10.962	2.597		13.854	3.974	

The frequency of GL citing varies from a minimum of 7,8% to a maximum of 62,9%.

The graph shows the variability of the frequency of GL citations for each source over the years



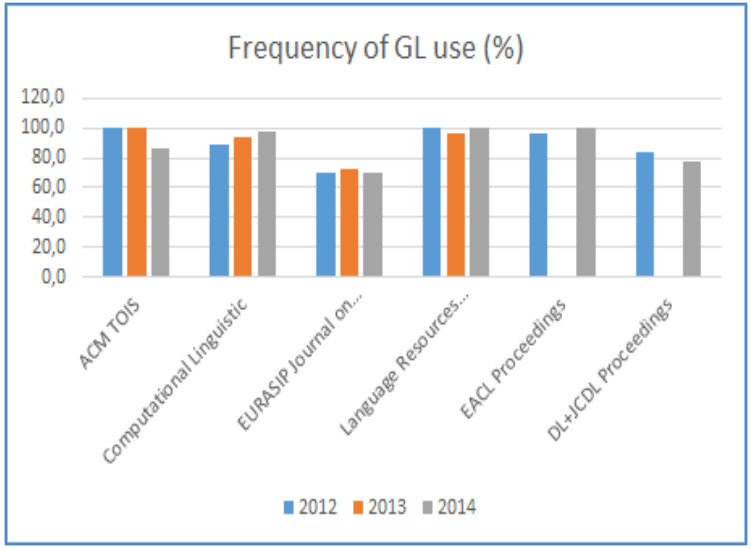
Analysis of data and results

Frequency of GL use

Frequency of GL use									
	2012			2013			2014		
Title	Number of articles	Number of articles with GL references	Frequency of GL use (%)	Number of articles	Number of articles with GL references	Frequency of GL use (%)	Number of articles	Number of articles with GL references	Frequency of GL use (%)
ACM TOIS	25	25	100,0	22	22	100,0	21	18	85,7
Computational Linguistic	36	32	88,9	35	33	94,3	34	33	97,1
EURASIP Journal on Advances in SP	252	176	69,8	188	136	72,3	183	128	69,9
Language Resources and Evaluation	31	31	100,0	56	54	96,4	31	31	100,0
EACL Proceedings	85	82	96,5	/	/	/	78	78	100,0
DL+JCDL Proceedings	96	81	84,4	/	/	/	97	75	77,3
Total	525	427		301	245		444	363	

In this table, the frequency of GL use shows the percentage of articles with at least one GL citation out of all the articles taken into account.

The frequency of GL use is globally very high and varies from a minimum of 69,8% to a maximum of 100%.



Analysis of data and results

Frequency/Intensity of GL use

Frequency/Intensity of GL use									
	2012			2013			2014		
Title	IF	Frequency of GL use (%)	Intensity of GL use (%)	IF	Frequency of GL use (%)	Intensity of GL use (%)	IF	Frequency of GL use (%)	Intensity of GL use (%)
ACM TOIS	1.070	100,0	20,2	1.300	100,0	12,3	1.021	85,7	16,0
Computational Linguistic	0.940	88,9	52,8	1.468	94,3	66,7	1.226	97,1	60,9
EURASIP Journal on Advances in SP	0.807	69,8	11,2	0.808	72,3	10,9	0.777	69,9	11,5
Language Resources and Evaluation	0.659	100,0	31,5	0.518	96,4	37,4	0.619	100,0	39,1
EACL Proceedings	/	96,5	39,7	/	/	/	/	100,0	56,3
DL+JCDL Proceedings	/	84,4	29,9	/	/	/	/	77,3	32,8

The overview of the frequency and use indicators related to the journals' IF doesn't allow to make considerations applicable to all journals.

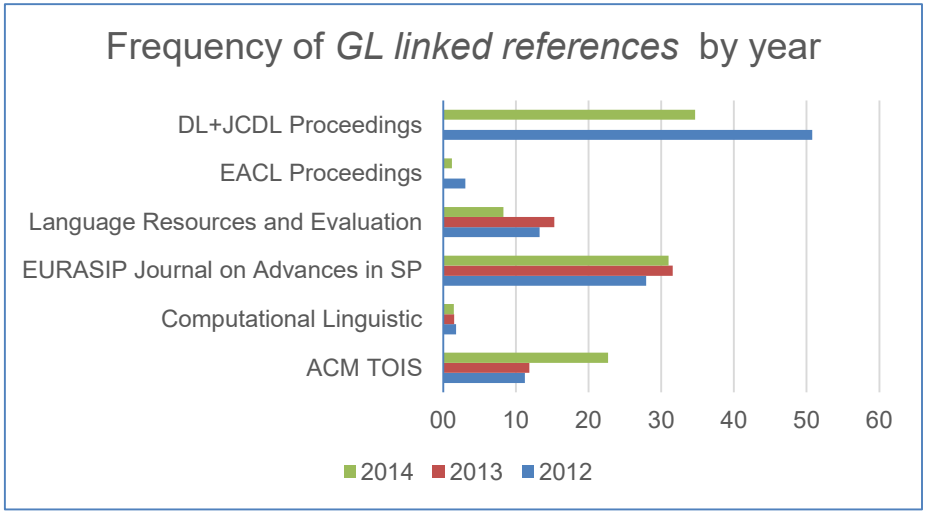
- In *ACM TOIS* the Impact Factor (IF) value seems to affect more the intensity of use than the frequency (of use): if the IF increases the intensity of use decreases as the frequency remains stable; conversely, if the IF of *Computational Linguistics* increases even the frequency and intensity of GL use increase;
- in *LR&E* the increase of IF seems to determine only the growth of frequency of use but not the intensity of GL use.
- the stability of the EURASIP IF seems to determine even the stability of all the indicators;

Analysis of data and results

Frequency of use of GL linked references

Frequency of the use of GL linked references									
	2012			2013			2014		
Title	Number of GL references	Number of *linked*-GL references	Frequency of *linked*-GL references (%)	Number of *linked*-GL references	Number of *linked*-GL references	Frequency of *linked*-GL references (%)	Number of *linked*-GL references	Number of *linked*-GL references	Frequency of *linked*-GL references (%)
ACM TOIS	285	32	11,2	135	16	11,9	150	34	22,7
Computational Linguistic	739	13	1,8	1.263	19	1,5	1.158	17	1,5
EURASIP Journal on Advances in SP	616	172	27,9	459	145	31,6	455	141	31,0
Language Resources and Evaluation	384	51	13,3	740	113	15,3	495	41	8,3
EACL Proceedings	884	27	3,1	/	/	/	1.332	16	1,2
DL+JCDL Proceedings	329	167	50,8	/	/	/	384	133	34,6
Total	3237	462		2597	2597		3974	382	

The frequency of use of GL linked references varies from a minimum of 1,5% to a maximum of 50,8%:



Analysis of data and results

- GL Typologies

2012																				
Title	Article	Blog/Forum	Book/Book chapter	Conference paper	Corpora	Course material	Database	Dataset	Deliverable	Guideline/ Normative document	Patent	Poster/ Presentation	Preprint/ Postprint	Report	Software	Technical Documentation	Tertiary document	Thesis	Undefined	Website
ACM TOIS	10	1	0	215	0	0	0	0	1	4	1	0	4	18	2	1	0	16	1	11
EURASIP Journal on Advances in SP	6	1	5	213	0	2	16	2	2	32	19	3	33	85	16	45	10	80	3	43
Computational Linguistic	1	0	3	632	0	2	0	0	0	0	0	0	4	36	2	6	13	38	1	1
Language Resources and Evaluation	1	0	1	291	0	0	2	0	0	1	0	1	4	35	9	8	3	21	0	7
DL+JCDL Proceedings	46	6	7	86	0	3	2	0	8	11	2	3	4	38	9	10	0	6	3	85
EACL Proceedings	0	2	3	771	1	4	2	0	0	2	0	1	10	46	3	3	6	24	0	6
2013																				
Title	Article	Blog/Forum	Book/Book chapter	Conference paper	Corpora	Course material	Database	Dataset	Deliverable	Guideline/ Normative document	Patent	Poster/ Presentation	Preprint/ Postprint	Report	Software	Technical Documentation	Tertiary document	Thesis	Undefined	Website
ACM TOIS	3	0	0	86	0	0	0	2	0	1	0	0	3	18	3	5	1	10	0	3
EURASIP Journal on Advances in SP	11	1	2	123	0	2	7	3	1	45	8	0	23	58	19	49	2	50	2	53
Computational Linguistic	4	0	0	1107	0	0	0	0	1	0	0	9	2	40	2	14	10	71	0	3
Language Resources and Evaluation	11	2	0	567	22	0	0	1	4	17	0	1	7	32	6	7	5	50	3	5
2014																				
Title	Article	Blog/Forum	Book/Book chapter	Conference paper	Corpora	Course material	Database	Dataset	Deliverable	Guideline/ Normative document	Patent	Poster/ Presentation	Preprint/ Postprint	Report	Software	Technical Documentation	Tertiary document	Thesis	Undefined	Website
ACM TOIS	4	0	0	97	0	0	1	0	0	5	0	1	4	15	1	1	5	9	2	5
EURASIP Journal on Advances in SP	1	2	0	113	0	2	21	8	5	22	9	0	21	71	11	56	3	60	4	46
Computational Linguistic	5	0	4	1031	1	2	1	0	0	4	0	0	4	38	1	11	7	44	2	3
Language Resources and Evaluation	2	2	0	402	2	6	2	0	3	5	0	2	7	26	2	3	1	30	0	0
DL+JCDL Proceedings	47	8	2	118	2	0	1	0	10	15	0	8	13	62	12	26	1	10	2	47
EACL Proceedings	15	1	2	1206	0	0	0		0	2	1	0	12	41	4	7	6	32	0	3

Analysis of data and results

The previous table reports the distribution of GL documents over document types. The most cited GL typologies are:

- Conference papers (highest number of GL citations),
- Reports,
- Thesis,
- Preprint/Postprint.

These four types of documents are the most cited, regardless of the year, the nature of the products analyzed and the area of knowledge to which they belong.

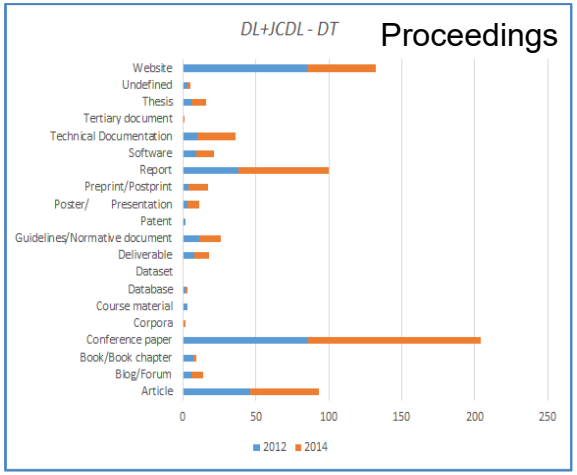
The type Article, although less frequently, is present in each resource and every year (*except for EACL 2012*).

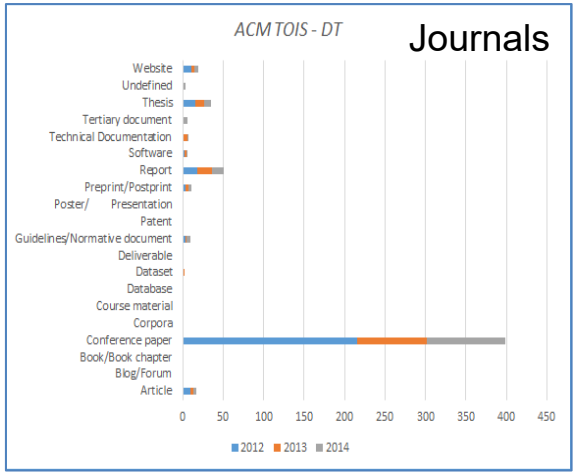
As for other types, some peculiarities related primarily to the topic of the selected journals and proceedings emerged:

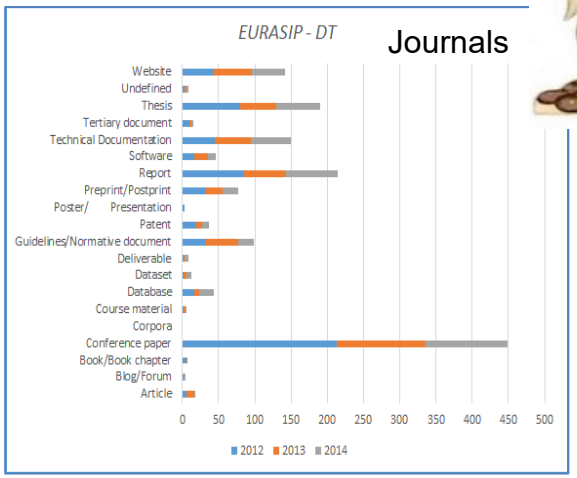
- Software/Tool, Technical Documentation, Database, Guideline/Normative document, Patent and Website are cited much more frequently in the E&E area than in Computer Science and Computational Linguistics areas.
- Corpora are cited much more frequently in the field of Computational Linguistics and, in particular, in the *LR&E* journal.
- None of the resources analyzed presents a considerable number of citations of Blog/Forum, Books/Book chapter, Dataset, Deliverable and Course material.

Analysis of data and results

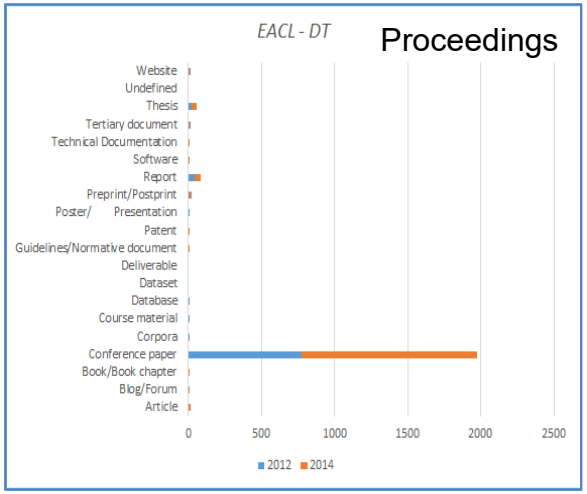
- *Document types in Computer Science and Engineering*

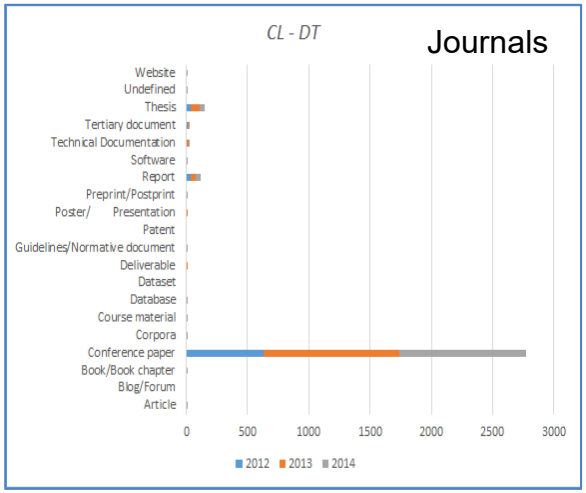


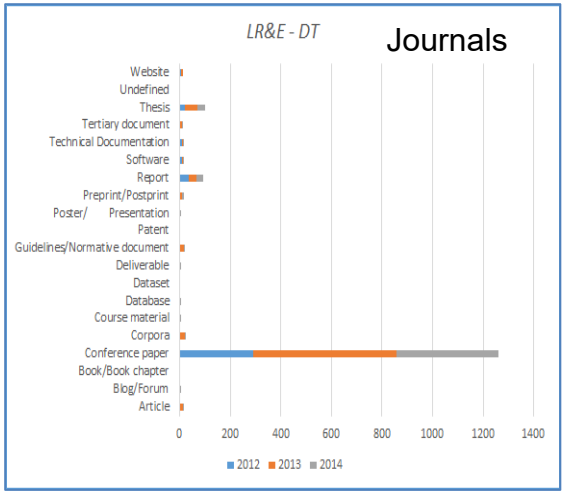




- *Document types in Computational Linguistics*







Conclusions

- In 2004 we analyzed two sample data belonging to two very different scientific fields.
  In this work the disciplinary boundaries of sample data are much less defined. Nevertheless, there are several significant differences, both in frequency and intensity of use of grey citations and about the cited type of documents, especially related to the specific field of study of each resource analyzed.

- The results obtained show that the Engineering domain has the least number of grey citations while the area of Computational Linguistics uses them most.
- *ACM TOIS* is the only resource comparable with data analyzed in our work of 2004: the analysis indeed shows that GL frequency of citing and use remains in the range of *11.5* to *21.1* values identified for 1995 and 2003.

- We can conclude by saying that the traditional citation model - i.e. the habit to cite mainly conventional literature - is still very strong and leaves little room for alternative models. However, this survey returns percentages of frequency and intensity in use of GL substantially important, especially in the field of CL.

Conclusions

It is increasingly clear the willingness of Associations and Organizations to collect, preserve and share the research results. The Repositories and the Open Access model have broken new ground and provided important tools for making these emerging communication needs come true.

Everything suggests, therefore, that the number of grey citations could increase in a very close future.

Thank you!