13th International Conference on Current Research Information Systems,
CRIS2016, 9-11 June 2016, Scotland, UK

The Italian model of distributed research information management systems: a case study

Paola Galimberti[a]*, Susanna Mornati[b]
[a]*Università degli Studi di Milano, Via Festa del Perdono 7, 20121 Milano, Italy*
[b]*Cineca, Via R. Sanzio 4, 20090 Segrate MI, Italy*

Abstract
The present contribution is divided into two parts. The first part, written by Susanna Mornati, describes the impressive progress of the CRIS adoption in Italy, brought forward as a personal experience leading the CRIS team. In less than one year, starting from September 2014 to July 2015, sixty-six Italian institutions (mostly universities and some research centres) adopted the same platform, a version of DSpace-CRIS called IRIS, customized for the Italian research environment. DSpace-CRIS provides an institutional Open-Access repository built in the solution: this was an unique opportunity to make publication data availbale on the Internet at a national level and gain insight on the Italian scientific production across several years. Moreover, the rapid adoption of ORCID at a national level gave the opportunity to enhance the quality of metadata.
The second part, written by Paola Galimberti, presents the results of her research on data extracted at a national level from the areas of Biology, Humanities and Social Sciences. For the first time in Italy these data were available in a standard then comparable format. The research ipothesis was to verify if and how the research assessment exercises of organizational structures (institutions and their departments) and of individuals had any influences on the publishing choices of Italian researchers. As a conclusion, the results are not univocal, and additional investigations are necessary for a deeper understanding of the phenomena taking place.

Keywords: CRIS; DSpace-CRIS; ORCID; research evaluation; publishing behaviour.

* Corresponding author. Tel.: +390250312945. *E-mail address*: paola.galimberti@unimi.it

1. Part one: context and pre-requirements

1.1. Introduction: the context

In 2004 in Italy a national assessment exercise (called VTR, more recently VQR) was introduced to evaluate the quality of research in universities and the main research centres. The Italian Ministry of Education and Research (MIUR) entrusted a public agency (ANVUR) with the responsibility for this activity. Since then, ANVUR decides and updates the evaluation criteria (mainly based on qualitative and quantitative parameters of research publications), collects and analyses data from participating institutions, assignes scores and publishes results. Based on these results, public funds for research are distributed to publicly funded institutions, for which the participation in the exercise is mandatory.

At the time, institutions did not have any information systems to sistematically manage data about research. When the campaign was repeated, in 2012, nearly half of the universities had adopted different research information management systems (RIMS, also called CRIS), mainly devoted to the collection and management of their research publications and interoperable with LoginMIUR, the national Ministerial closed database where publications are collected and referenced for grant applications.

In 2012, the first most diffused platform was UGOV-Research (adopted by nearly 40 institutions), a closed system developed by Cineca, the main Italian interuniversity consortium. The second was SURplus (adopted by nearly 10 institutions), an open OAI-PMH system developed by Cilea, the second Italian interuniversity consortium.

Still in 2014 the situation related to the management of data on research in Italy presented a number of critical issues: the most diffused system was closed and data were not publicly available, the Ministerial system was also closed, its data only available to individual/institutional owners, duplicate records proliferated, metadata were poor and limited, no validation processes were in place to check accuracy and consistency of the information entered.

Nonetheless, this mass of unverified data was still used to analyze and evaluate the quality of scientific production and the performance of institutions. In fact, the issues related to the poor quality of data, inconsistency, lack of transparency, emerged when ANVUR decided to use data in the ministerial closed database for a number of activities related to research evaluation: the VQR national exercise, as mentioned before, but also the evaluation of Research

Departments, the national advancement of researchers' and lecturers' careers, the ranking of Italian scientific journals, and so on. Some of these issues have also been echoed in the international press [1].

The unification of the Italian interuniversity consortia in 2013 provided a framework for Italian universities to suggest a collaborative effort that led in 2014 to the release of a new CRIS, called IRIS [2], based on DSpace-CRIS, which merged and enhanced the previous systems, providing a range of useful tools for system management and related procedures, such as workflows for quality assurance, copyright management, and a repository for open access publications. As an extreme challenge, 55 institutional systems were migrated by Cineca from previous CRIS platforms to IRIS, mostly in a tight timeframe from September 2014 to July 2015, and 11 new installations were released, for a total amount of 66 IRIS instances (a full list with URLs for public access is available here: http://tinyurl.com/h73c3n2). The success of the huge operation was guaranteed by the strong competence and experience of the project team, the technical excellence of the solution, the high-level change management strategies that were adopted.

1.2. Benefits of a country-wide adoption of the CRIS

DSpace-CRIS (https://wiki.duraspace.org/display/DSPACECRIS/DSpace-CRIS+Home) is a free open-source interoperable CRIS platform that collects, manages, preserves and disseminates information related to the entities that populate the research domain, such as resources (e.g. researchers and their research interests, references, positions, organizational units, laboratories, research groups, etc.), activities (e.g. projects, contracts, collaborations, etc.), outputs (e.g. publications, reports, patents, etc.). Entities with their attributes and their relations are modelled in compliancy with the CERIF ontology (http://eurocris.org/cerif/main-features-cerif), which is the de-facto world standard for the modellization of the research information domain. DSpace-CRIS implements workflows such as submission, validation, publication of items. It allows import/export and interoperability of data and processes with external databases and applications, for instance bibliographic and bibliometric information from Scopus and Web of Science, PubMed, CrossRef, etc. It allows the assignment of DOIs for datasets via interactive transactions with Datacite. It is compliant with EU requirements for Open Access via OAI-PMH exposure for OpenAIRE. It provides webmetrics for item views and downloads, and aggregated metrics for authors.

DSpace-CRIS SEO features make it a very powerful tool to enhance visibility and impact of researchers and their institutions [3]. Its Italian version, IRIS, allows uploading of publications to the Ministerial National Database, LoginMIUR, and provides other local procedures such as procedures for qualitative evaluation and quantitative reporting that are relevant to the Italian community. While DSpace-CRIS is open for international collaboration, IRIS has an institutional focus group that was established by Cineca in 2015 to suggest evolutionary requirements of the platform and decide on priorities in his development. The group is formed by representatives of Italian universities and research centres that adopted IRIS, with mixed technical profiles (CIOs and IT staff, research support officers, librarians, etc.).

In May 2016 IRIS installations sum up to 3,8 million publications.

- 3,5 million of these are publicly available on the Internet (nearly 10% with open-access full text);
- 2,8 million items are harvested by PLEIADI, the Italian portal for scholarly literature in open archives and institutional repositories (http://find.openarchives.it/);
- over 6 million records are input or uploaded in the closed Ministerial database LoginMIUR (the higher number is due to the duplicates produced by coauthors who have to populate their individual space).

Tha availability of this huge amount of data enhances the knowledge of Italian scholarly activity, the quality of data and the transparency of evaluation procedures.

1.3. The "persistent identifier" approach

ORCID (https://orcid.org/) provides unique persistent identifiers for researchers. It provides disambiguation of authors name, tracking of name variations, and allows scholarly authors to gain full credit for their work across all systems integrated with the ORCID registry. It enables transparent and trustworthy connections between researchers, their contributions, and affiliations, helping people find relevant and accurate information and simplifying reporting and analysis.

In April 2015 the first CRIS/ORCID integration for DSpace-CRIS was released. In May ANVUR announced that ORCID would be mandatory as the researcher identifier for the next VQR (2011-2014), starting in Autumn 2015, and a consortium agreement was signed to provide ORCID services to all Italian institutions. In June Cineca released an open-source ORCID Hub, a national central register of IDs and authorizations, to provide easy coordination of the ORCID effort at the national level, registering all ORCIDs related to affiliates to Italian universities and research institutions and their authorizations to data interoperability with local applications and the Ministerial database. The Hub is also useful for organizations that do not have a CRIS, to keep trace of the ORCID adoption rate by their affiliates [4].

By November 2015 (less than 6 months after the first announcement of the project) the vast majority of Italian researchers (over 80%, representing nearly 60,000 researchers in 60 different institutions) was identified by an ORCID [5].

The adoption of ORCID in a wide spatial context (national) and rapid temporal context (a few months for implementation) is a first and unique example in the world. The Italian ORCID Hub constitutes an original and interesting case study for research infrastructure and it deserves to be carefully observed also in the future. The narrative and the results of the Jisc ARMA ORCID project [6] in UK were very helpful, but the singularity of the Italian case was the speed of the adoption at national level.

The availabilty of ORCID at a national level improves data quality, interoperability and constitutes the first building block for a robust data infrastructure based on the unambiguous identification of all entities that populate the research domain.

1.4. Lessons learned

The adoption of the same CRIS system is only the first step towards stronger harmonization of procedures. The tool currently used by 66 universities and research centers is a work in progress that needs to take account of developments in the field of scholarly communication and data management.

The availabilty of an institutional repository allowed the adoption of an Open-Access policy by most institutions, raising awareness about the potential impact and visibility of Italian research. The analysis of data from the Italian CRIS installations shows that the design of CRIS scales nationwide and thus makes it possible to compare data among a significant number of instances.

The nationwide adoption of ORCID and its harmonization in all CRIS installations will make it possible to adopt an architecture of data at the level of the links between authors and publications, publications and research projects, projects and investigators, and so on. Disambiguation of researchers' names will allow an even more extensive interoperability worldwide. So data from the local systems can be harvested for the implementation of a national open disambiguated database of the Italian research.

A relevant finding concerns the experience of sharing the domain model and the architectural and functional needs among experts from different institutions. The opportunity to operate within a focus group that determines the evolution of the system provided an important boost to the development of the CRIS. To share the same tool, collaboratively developed by a consortium of universities, also allows to collaborate in improving data quality, sharing records and reduce duplication of work.

The enormous amount of data, open, comparable and interoperable at a national level, offers for the first time in Italy the possibility to investigate and understand publication strategies, strengths and weaknesses, main research areas, emergent areas, a picture of the Italian research (including Humanities and Social Sciences) until now not imaginable.

Italian data about research publications in CRIS are quite complete for the most recent years. It is mandatory to collect data about research output in order to participate in the national research assessment exercises and research funding programs. However, data quality is still not uniform and not all institutions care about quality assurance (lack of awareness) and implement a validation process of the bibliographic information submitted by researchers, so this may negatively influence data analysis, particularly at the meso and micro level.

2. Part two : data analysis

2.1. The evaluation of research in Italy

In recent years in many European countries research funding systems based on performance were introduced (the so-called Performance-Based Funding Systems, PBFS) [7]. The object of the introduction of these systems is to ensure, in an era of scarce resources, that the best research is funded. The idea is also that through the introduction of these systems virtuous mechanisms can onset that lead to an overall improvement of the system. The assessment of the effectiveness of these systems is not univocal as it is not yet clear whether the introduction of PB systems lead to a real improvement of the research or rather only to an improvement of indicators. It is feared that in many cases the indicators become the purpose and not the means for improving the quality of research.

Over the past 16 years Italy has seen three national assessment exercises on research. The first was VTR 2001-2003, the second VQR 2004-2010, and the third VQR 2011-2014 (nearing completion). In particular with the second and third evaluation exercises Italy has developed a performance-based system whose outcomes and whose unintended effects are not entirely clear to date.

The first Italian national assessment exercise had only limited financial consequences, but starting from 2013, and as a result of the second national assessment exercise, funding allocated on the basis of performance criteria rose in Italy from 7% (2009) to 13.5% (2013), with a further increase in the following years.

The National Evaluation Agency (ANVUR) performed the national exercise on the basis of a mixed system, peer review and bibliometrics. Both methodologies aroused a debate in Italy about the unintended consequences of the introduction of this funding system.

The "stability law" on budget allocation for 2015 stated that 18% of annual funding would be allocated to better performing institutions according these criteria: 70% on the basis of the results of the VQR, 20% on the basis of the scientific production of professors promoted or recruited in the period under assessment, 10% on the basis international teaching activities.

Although the evaluation exercise (2004-2010) carried out by ANVUR is very big as far as the absolute number of publications submitted is concerned (195,000 papers from 130 institutions), it concerns only a small part of the research carried out in the period under assessment, specifically 3 publications for each professor and researcher recruited before 2008.

2.2. Methodology

In order to understand if and how the ongoing evaluation systems (VQR but also other campaigns such as the national evaluation of researchers for career advancement, called ASN) influenced the editorial choices of Italian researchers, we analyzed the entire scientific production in the period from 2009 to 2014, since it was the first time that a great amount of data were available, supplied by the Italian universities that adopted the new CRIS.

The period chosen is the one that crosses the two latest assessment exercises, ranging from 2009 to 2014. We chose 2009 as the basis of comparison because in that year the PB evaluation system had not yet been introduced, using data from 2009 and 2010 as a point of comparison. The areas considered are the ministerial areas 05 (Biology), 10 (Philology, Literatures and Arts), 11 (Philosophy, Psychology and Pedagogy), 12 (Law), 14 (Political and Social Sciences). For the area 05 we examined publications with Impact Factor™ (IF) and IF location in percentiles, for HSS areas the types of publications and the language were considered.

In the case of Biology we tried to figure out if the emphasis of the journal IF as one of the evaluation criteria used did increase publications in journals with IF and especially if it has increased the number of publications in higher quintiles. For Humanities and Social Sciences (HSS) areas we tried to figure out if the value given to journals listed in A-category (considered as best quality) has oriented researchers more towards this type of publications thus abandoning other typical products of these scholarly disciplines such as the monograph and the essay.

We are aware that not all data are of excellent quality, nor has it been possible to eliminate duplicates due to the fact that the records of the various archives are added together and in the case of co-authors of several universities the same record is repeated several times, as well as account has been taken of the fact that if a researcher moves from one university to another they duplicate their publications in repository of his new institution. However, we estimate that this noise did not undermine the significance of the results obtained.

2.3. Analysis

For the area 05 125,289 records were considered. They consisted of journal articles, including 71,008 articles with IF and 54,281 articles without IF. In the time period examined, the number of articles in journals with IF remains more or less constant, while that of the articles in journals without IF decreases [Fig. 1].

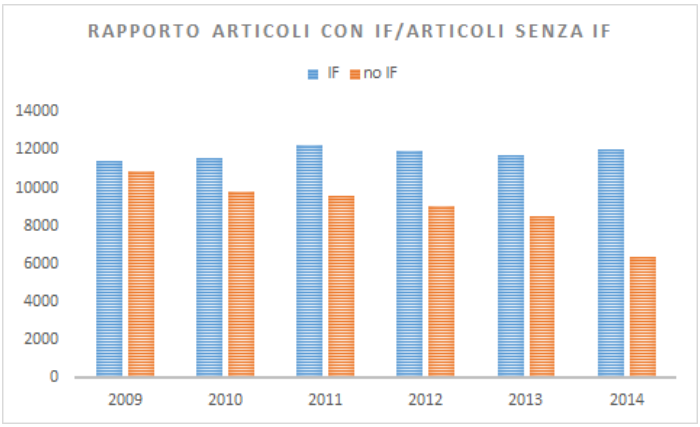

Fig. 1, Period 2009-2014: ratio of articles in journal with and without IF

The reduction of number of articles in journals without IF however, cannot alone be explained by the introduction of a funding system based on performance. To see if this system has steered researchers toward higher quality publications (not only those with IF) it is necessary to analyze the number of articles in different IF value ranges [Fig. 2].

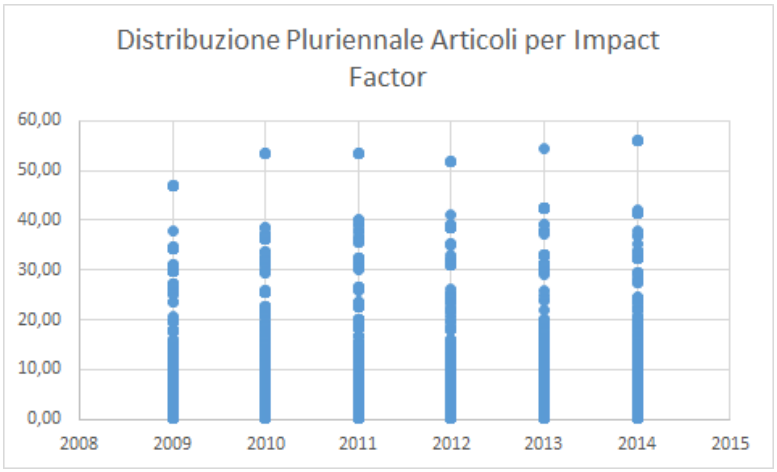

Fig. 2, Area 05 (Biology): Distribution of articles in different IF value ranges

An analysis of the distribution of IF value ranges shows that most of the articles concentrate in the 0-8 range, with indeed very few publications in the higher ranges (from 8 IF upwards). This situation does not change in the time period under review (2009-2014).

To these data we added a further analysis based on the publications of the area 05 for the 2011-2014 period, which shows an improvement for at least the average figure. The criterion used was that of Field Weighted Citation Impact™ (FWCI) borrowed from Snowball metrics [8] and analyzed through the SciVal database [Fig. 3]. FWCI divides the number of citations received by a publication by the average number of citations received by publications in the same field, of the same type, and published in the same year. The world average is indexed to a value of 1.00. Values above 1.00 indicate above-average citation impact, and values below 1.00 likewise indicate below-average citation impact.

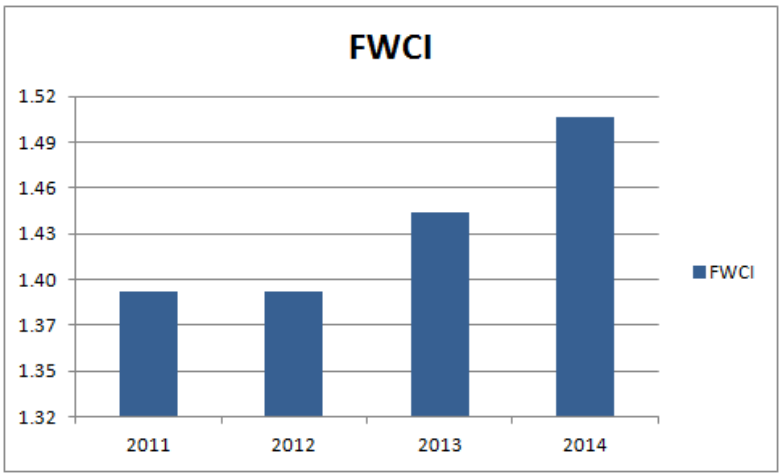

Fig. 3, Area 05: FWCI in period 2011-2014

To verify the variation in the percentage of publications in the top 5%, 10% and in the top 25% worldwide, we used SciVal, although in perspective the same data will be available for extraction from the local CRIS thanks to the import of percentiles data through the premium Scopus APIs [Tables 1, 2, 3].

year	output % in top 5%
2009	9,6
2010	9,6
2011	10,1
2012	9,9
2013	11,1
2014	10,6

Table 1. Percentage of publications in the top 5% worldwide (self citations excluded)

year	outputs in top 10%
2009	17,2
2010	17,9
2011	19
2012	18,5
2013	20

2014	20,6

Table 2. Percentage of publications in the top 10% worldwide (self citations excluded)

year	outputs in top 25%
2009	38,2
2010	41,7
2011	43,1
2012	43,7
2013	40,3
2014	48,4

Table 3. Percentage of publications in the top 25% worldwide (self citations excluded)

Bibliometric data on publications seems to confirm previously illustrated data on journals, for which the number of excellent publications remains constant (top 5% and 10%) while the number of items that are placed in the top 25% grows (among publications of the same year, the same type and same subject area). The analysis shows that in the biological sciences there has been a moderate effect by the introduction of a PBFS, probably mitigated by the issue of the S. Francisco Declaration on Research Assessment (DORA, http://www.ascb.org/dora/) that was mostly supported by biological scientists. Further investigation should be done in areas like Chemistry or Medicine to see if the influence of the new research funding system on publication behaviour is confirmed.

As for the areas of human and social sciences, the dataset includes 48,572 records examined in the area 10, 54,150 records for the area 11, 52,799 records for the area 12, 17,787 records for the area 14. For these areas, we investigated to what extent the regulations for the national evaluation of researchers for career advancement (ASN) and the VQR influenced publishing strategies in HSS, possibly fueling opportunistic behavior.

The ASN laid the rules (not binding but used for many disciplines) to get access to the qualitative assessment procedure that involved the overcoming of different thresholds for each competitive sector (medians). What was immediately observable was the huge disproportion between the number of monographs and number of articles in the A-range journals, necessary for a nomination to be considered for evaluation (in some areas publication of just one article in any A-range journals was enough to access the procedure) [1].

Another issue is that of internationalization of research: from the 2004-2010 VQR internationalization emerged as one of the criteria for the qualitative evaluation. Even in HSS areas publishing in English language began to be considered a virtue. The analysis of data at a national level is meant to verify what kind of effect these criteria had on publication strategies of Italian researchers in HSS disciplines [Tables 4, 5, 6, 7].

Year	articles	book chapters	monographs
2009	2248	5021	713
2010	2208	5378	699
2011	2364	5438	699
2012	2376	5753	956
2013	2385	4358	597

2014	2286	4617	469

Table 4. Distribution of publications by type in area 10

Year	articles	book chapters	monographs
2009	3615	4672	944
2010	3641	4525	825
2011	3760	5117	856
2012	3899	4653	1046
2013	4114	3751	676
2014	4147	3342	560

Table 5. Distribution of publications by type in area 11

year	articles	book chapters	monographs
2009	3195	5215	644
2010	3013	5414	631
2011	3112	5615	629
2012	3122	5489	1200
2013	2867	4404	660
2014	2862	4297	423

Table 6. Distribution of publications by type in area 12

year	articles	book chapters	monographs
2009	862	1743	354
2010	849	1730	356
2011	956	1831	316
2012	958	2002	477
2013	933	1661	278
2014	892	1331	251

Table 7. Distribution of publications by type in area 14

The data for the areas examined are not univocal. While there is a clear decrease in the number of monographs (works that certainly take a long production time and thus not highly scoring for evaluation purposes), we notice a slight decrease in the number of book chapters, also works that require more processing time, except in Area 12 (Law) where

the number of items remains constant or increases slightly. This leads us to think that probably the researchers prefer to choose less timeconsuming formats [Tables 8, 9, 10, 11].

year	eng	fre	ger	ita
2009	1008	169	143	4908
2010	1061	166	161	4784
2011	1152	162	140	4874
2012	1313	181	128	5189
2013	1182	117	124	4219
2014	1338	120	119	4158

Table 8. Distribution of publications by language in area 10

Year	eng	fre	ger	ita
2009	1369	133	73	5917
2010	1581	120	73	5253
2011	1815	70	69	5618
2012	2061	69	59	5359
2013	2350	58	53	4663
2014	2578	67	56	4058

Table 9. Distribution of publications by language in area 11

year	eng	fre	ger	ita
2009	425	50	26	6833
2010	454	40	40	6676
2011	539	34	32	6856
2012	636	35	52	6867
2013	641	27	26	5652
2014	632	30	30	5510

Table 10. Distribution of publications by language in area 12

year	eng	ger	fre	ita

2009	406	29	12	2015
2010	413	19	13	1908
2011	478	19	18	1938
2012	548	26	26	2221
2013	576	19	18	1822
2014	588	11	9	1500

Table 11. Distribution of publications by language in area 14

The figure on the language appears more uniform in the various areas, though their “lingua franca” is different. This is not the seat for a discussion on the meaning and the importance of national languages and the variety of “lingue franche” related to the various disciplines. Here we will just notice a decrease in contributions written in Italian and an increase in those written in English. This is reasonably due to the importance given to the use of English as the language of communication of science regardless of disciplinary specificity.

Conclusions

An analysis based on such a vast amount of data from CRIS had never been previously done in Italy. This study was made possible by the adoption of the same CRIS platform in the course of 2015.

In conclusion, this first analysis of data showed that for the area of Life Sciences the introduction of reward criteria in 2004-2010 VQR, also based on the IF, has not substantially changed the publishing habits of researchers in this discipline compared to the number of contributions and products and compared to the "quality" of the journals chosen.

On the contrary, we observe a decrease of articles in journals without IF. Further investigations should be carried out in other areas of Science, Technology and Medicine (STM) to understand whether these strategies are diffused or whether in the Life Sciences there has been an influence by the movement that led to the publication of the S. Francisco Declaration.

As regards the HSS, our investigation on the choice of channels of communications (publication types) gave different results in the various areas examined, even if we see a general attitude more inclined towards the publication of articles rather than of monographs or book chapters. This attitude could be caused by the admission criteria to the ASN, which required very low thresholds for the type of articles in A-range journals , therefore easier to overcome. Also in this case, further investigation should be made with respect to the quantity of articles in A-range journals, compared to articles published in scientific journals.

As for the language used, contributions in English have increased in all areas analyzed, even in the notoriously conservative Law, in response to the criteria used for the peer review in VQR 2004-2010, where the internationalization criteria has often been confused with the mere use of English language.

While we are relatively certain that this data quite faithfully reflect the Italian scientific production (where the bibliometric databases cannot due to poor coverage of HSS production), we are aware that data quality can be improved. Italian universities are collaborating to achieve greater uniformity in the quality of the data. It is our interest to develop further research also in comparison with data from other countries.

References

1. Mazzotti M. Listing wildly. THE 2012;Nov.8th. https://www.timeshighereducation.com/features/listing-wildly/421723.article?storycode=421723
2. Bollini A, Mennielli M, Mornati S, Palmer DT. IRIS: Supporting & Managing the Research Life-Cycle. UJER 2016;4(4):738-743. DOI: 10.13189/ujer.2016.040410
3. Palmer DT, Bollini A, Mornati S, Mennielli M. DSpace-CRIS@HKU:Achieving visibility with a CERIF compliant open source system. J ProCS 2014;33:118-123. DOI: 10.1016/j.procs.2014.06.019
4. Bollini A, Brown J, Mennielli M. Think global, act local: innovating nationally with the Italian ORCID hub. Poster presented at Force2016, Portland, OR, 2016 April 17-19.

https://www.force11.org/entityform/104
Awarded: https://www.force11.org/meetings/force2016/force2016-best-poster-and-demo-winners
5. ANVUR (2015) Progetto IRIDE
http://www.anvur.org/index.php?option=com_content&view=article&id=829&Itemid=600&lang=it
6. Jisc ARMA ORcid Project https://orcidpilot.jiscinvolve.org/wp/
7. Jonkers K., Zacharewicz T. (2015) Performance based funding: a comparative assessment of their use and nature in EU Member States
http://publications.jrc.ec.europa.eu/repository/bitstream/JRC97684/jrc97684_pbf%20final%20.pdf
8. Snowball metrics https://www.snowballmetrics.com/