6e Séminaire SciScI - OST
Vendredi 12 octobre 2012
Isidro F. Aguillo (CSIC, Cybermetrics Lab), Espagne

"Information metrics (iMetrics): New Developments"

TRANSCRIPTION

First of all, sorry for speaking in English. Unfortunately my French is not good enough for such a various audience, but I'm also taking into account that today is the Spanish National day, so it's a bit strange for me, as a Spaniard, to celebrate our National day in Paris, but you need to think about that. So it is a really crazy situation for me today.

Second point, of course: thank you to OST for inviting me to this presentation, all of you for attending. At the end of the presentation, I'll invite you to discuss and exchange ideas about the topics I'm going to introduce to you today, because I think what is as important as the information I intend to show is to learn about your opinion, your experience or your criticism about the work we are doing. We think this kind of exchange and your input are very important.

The presentation will be available on OST's webpage. You can download it and use it as you prefer, since one of the most important things I'm supporting is opening up to initiatives and open access. So feel free to use the slides, but please cite in the source and the author.

This is a very personal presentation. It is our individual point of view on theses topics. It is not complete – since we only have three quarters of an hour to explain our work. There are many other tools that can be included, but I think this overall presentation will give you a good idea of the topics.

You know that, even today, bibliometrics is playing a very important role. Yesterday I had a meeting with the people at OST and of course bibliometrics and the traditional

citation databases are still very important. You know that the situation has been changing, even in the field of bibliometrics, in the recent years, especially because now we have relevant new citation databases. I'm talking about the free citation databases. The word "free" is capital here, because you know in Europe we have a general funding crisis and the access fees on traditional citation databases are very expensive. In the future, the only alternative might be Ñ not because of technical reasons, but for economical reasons Ñ a large free database. Of course you can also consider the new generation of bibliometric indicators that have been developed in the last years, especially with the introduction of the h-index. There are a lot of new developments there, but we are not going to talk about them in this presentation.

There are still many problems regarding bibliometric analysis. There is a maturity in the area, but problems remain. Perhaps one of the most important is the freshness of data: now everybody wants to know the current situation, for instance the impact of the crisis on the research and development arena. The big problem is the data we are collecting is generally two or three years old, so it's not easy to understand the current situation, taking into account the information is so old. We need a citation analytic window that is less than one year old.

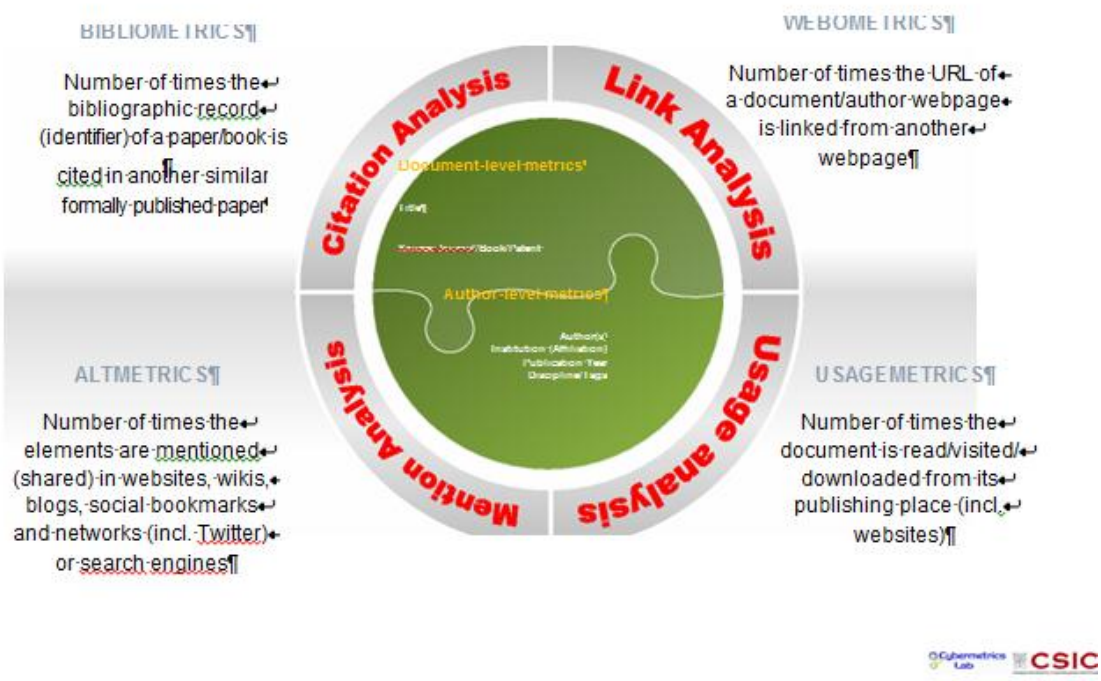
The other thing is bibliometric data is regarding peers, regarding only a close environment formed by the academics, university professors, scholars, scientists, and perhaps Ñnot only for innovation purposes, but also for the impact on society Ñ we need to know not only the impact on this small group of actors, but also the impact on the rest of society. Non-academic sectors are very important as additional stakeholders, and as you know citation analysis is only concerning the links between partners inside Academia.

Finally, perhaps we need more information if we intend to expand this universe of actors to the whole society and to better explain the motivations behind impact and visibility of scientific activities. We need to build a better theory that explains not only the resource for traditional scientific citations, but also non-scientific linking in the environment of Academia and research centers.

We are focusing this presentation on the emerging iMetrics sub-disciplines. Some of them are already a bit old and started in the middle or late nineties, but other new sub-disciplines are really very young, only two or three years old, so it can be very interesting for you to know what the current developments are in these new emerging disciplines.

Their main purpose is to explore larger and more diverse audiences, even in the scientific field Ñ for example researchers coming from developing countries. I have worked with Latin-American colleagues who often said: “We are playing an important part in the development of science in our country, that is very important for our economy, but we are isolated from the rest of the world”. This clearly needs to change. You need to take into consideration not only your peers in Western countries, but also the new actors in Asia, in Latin America, even in Africa. This, I think, is crucial.

Bibliometrics currently focus on formal communication (papers, publications in international or high-impact journals). We are now recognizing the role and impact of the informal communication, also including some kinds of new processes. We need to combine all of them, and to build a new portfolio of indicators.



A New model for the XXIst century

When you are using bibliometrics, you have basically two main indicators: the number of papers, and the number of citations of these scientific productions. You can of course combine both, and you can use other indicators, but basically the current portfolio only consists of two or three different indicators. We need to enlarge this portfolio to incorporate new alternatives.

Here is the key of my presentation. This is our current view of the iMetrics for the new century.

You have traditional bibliometrics, using citation analysis Ñ everybody knows about this discipline Ñ more or less build on statistics and very useful, especially in Europe, whereas in the U.S. the situation of bibliometrics is different and the discipline is not so popular. You have webometrics, which we are going to talk about and is part of the emerging sub-disciplines I mentioned earlier. Link analysis is already fifteen years old, it started in the mid-nineties, so we have a lot of information. Altmetrics is a completely new development, since the manifesto is only two or three years old, and it is based on mention analysis. You can see we are naming each discipline as well as the most commonly used analytical tool for each. We need to focus on the presence of scientific information in the current new web-communication tools.

Finally, usagemetrics analyses the consumption of scientific information that is available on the web. It is so new that there are really few papers or contributions in this area. There are still technical problems, but it could be one of the most promising disciplines. Here we are talking about possibly hundreds of variables involved. If we have access to the academic websites, the amount and the diversity of information is really very promising for developing studies in the future

I'm going to focus on a new development of bibliometrics Ñ clearly bibliometrics is not the core of this presentation, but I think it is important to consider one part of its development that is very related to the web-environment. I'm talking about the profiling. I'm going to introduce you the two new citation databases I mentioned before: Google Scholar and Microsoft Academic Search. One of their common characteristics is that they are both free to use. You can go on the web, scan your name or the name of an institution and obtain a lot of data. Google Scholar is far larger than any other competitors: WoS and Scopus have approximately 40 million records, whereas Scholar has over one hundred million records. Academic Search is based on similar databases

to WoS or Scopus so their number of records is really close. Academic Search data may be more reliable, as it is closer to the data source than what was accepted for WoS and Scopus.

Google Scholar is collecting data from far different sources. Perhaps, individually, every source is reliable, but when you are combining them, there is a serious problem of quality control. The big advantage of Google' s tool, apart from its large size, is the current information. They are really adding information every day, so you can check information from two weeks ago and compare it with yesterday's. Unfortunately, it is not the case with Academic Search. I think there will soon be a better version, but the problem is the information they are offering now is basically the same as last summer, when the database was introduced. So it is already a year old, and you need to consider that.

But this is not where the main interest of these products resides for me. For me, the most interesting is the two mechanisms, the two profiling systems those two giants developed. In the case of Google Scholar Citations, it's a very interesting system because it is made on a voluntary basis: you need to register in order to create a profile that is automatically populated by Google Scholar. You can personalize your profile, correct and update information about your work, etc. With MS Academic Search, the system is totally different. They automatically generate the profiles, and then they offer you the possibility, if you register, to tune up your record. The total number of profiles in Google Scholar Citations is about one hundred thousand researchers from all over the world, but the total corpus of Academic Search is 20 million. You need to take into account that many of these 20 million are duplicates Ñ they don't have enough people to clean up their data from the duplicates and unify the profiles.

Here is an example. This is the profile of a very famous infometrics researcher from Israel. She decided to create herself a profile. She needed to register, and she obtained this page. You find information about her published papers in there, but also the number of citations, and the number of co-authors. You must consider that the presence of the photo and the names of the disciplines were added by the author herself. In this case, you can see she doesn't have webometrics in her specialties. This is a decision she made. It could be problematic because she cannot be found by using the keyword search "webometrics". As you can see, the topics are in blue, which means they are links, and if you click on the links you can access a ranking on this topic according to the data available under Scholar Citations. This is very useful, not only because of the ranking

you get, that is of course never complete. The interesting thing for me, after long debates about new bibliometric indicators, Google Scholar very cleverly decided the best indicator for ranking people is the good old-fashioned number of citations. No H index, no i-10 index, no sophisticated index, but the traditional total number of citations.

This is the MS Academic Search alternative. As I said, this is made automatically. In this case, the researcher has three different profiles. This one he probably made himself, he doesn't have time to combine it with the other ones, so there is a lot of noise here, but you can see the presentation is richer than on Google Scholar Citations. You have graphs, information about the co-authors, the topics, the journals he was published in, and so on. Very important also: the to pics are not coined by the authors themselves, but automatically provided by the system. This approach is therefore similar, but not equal to Google's.

The possibilities offered by the visualization of data are very interesting too. This too is given automatically, and it shows the relationship between our researcher and other scientists. You can doubt the quality of this information, but many other visualization tools are available. Obviously, if everybody is adding photographs, at the end we could have a pretty efficient database.

The possibilities offered by the visualization of data are very interesting too. This too is given automatically, and it shows the relationship between our researcher and other scientists. You can doubt the quality of this information, but many other visualization tools are available. Obviously, if everybody is adding photographs, at the end we could have a pretty efficient database.

One of the additional advantages of MS is they are providing profiles not only for authors, but also for other units, like institutions. The approach is similar to the one we registered in the case of individual authors. There are not many institutions for now... I looked for OST, and it is not present in the database. But as I said, it is a Beta version.

Now I'm going to talk about the new discipline of webometrics. I've been working in this discipline for fifteen years, and the situation changed completely during that period. At the turn of the century, webometrics were interesting but not central to study academic communication, whereas today, without any doubt, I think the web is already the main scholarly communication tool, and everybody should be aware of that. This means a lot of advantages: we are considering huge audiences Ñ the total number of people using the Internet in the world is over two billion. The information provided is very rich and

diverse (not only formal papers in peer-reviewed journals) Ñ even coming from the Academia, you've got very important informal information (blogs, social networks, loads of webpages...) even Wikipedia has become useful to reference topics Ñ and with many additional actors involved, and this is relevant for OST, I think. It takes into account not only the close environment of scientists, but also people benefiting, or using, or exploiting the results of researc h (economical and industrial sectors for instance), but also the implications for policy, the community engagement, not only the technological transfer, but the knowledge transfer. You need to know the people on the other side of this process. This is one of the main advantages of the web approa ch. You have access to these people, you have access to the true impact your actions are having on these people, or at least their opinion about the research you are doing.

This is not a revolution. The term "webometrics" was originally coined with an "o" by Professor Ingwersen from Danemark, because he wanted to state the close relationship between this discipline and bibli-o-metrics. This seems very important to me. When people say "webmetrics" and "webometrics" are the same, I can't really agree: the basic idea here is we are going to try and use webometrics as a complementary tool to bibliometrics, not instead of bibliometrics, but jointly to bibliometrics. That means also using the same ideas, and sometimes tools similar to those that are standards in bibliometrics and scientometrics.

We have powerful advantages. The most important is we can consider as a source commercial search engines like Google, Yahoo or Bing. We are talking about public databases in the order of eight hundred thousand million webpages, billions of links only in Europe... Huge databases. The main tool is link analysis, which is an extended and expanded citation analysis. From this perspective, link analysis is richer, more diverse, but that also means it is more problematic, there is more noise, data and motivations are more difficult to interpret. You have a freshness contents Ñ Google is updated almost instantaneously, every minute. For example, the most important newspapers are including data about the last scientific Nobel Prize at the very moment it is announced in Stockholm. From the point of view of authors and content producers it is a universal tool that is very easy to edit, cheap to access Ñ you don't need to be from a big and very rich western university or lab to produce an interesting webpage, you can contribute even if you work with a small agricultural laboratory in Indonesia, for instance Ñ and this is an example I know.

There are also major disadvantages, of course. I'm not trying to sell you only the advantages; we also need to know about the negative sides. There is a problem related to the theoretical background. Of course we need a theory related to the network organization of the academic world, we need a lot of epistemological developments... Of course here we are totally out of my experience and specialty, but there are problems related to citation, linking behavior and motivations, and the lack of articulation of the network theory in STS (Science, Technology & Society).

On the practical side, there are also a lot of problems. The major one is that the sources we are using (i.e. public search engines) are really opaque and therefore not easy to understand and crosscheck. They are based on compilation, indexing and ranking algorithms that are non-disclosed commercial secrets. What is the value of Google? Well, it is the secret formula of their ranking mechanism. It's like the formula of the Coca-Cola. Without it, Coca-Cola is nothing and the market value of Coca-Cola disappears. It is probably the same for Google, although for us, from a technical point of view, it is problematic.

Another problem is that we have several sources, and we discovered their geographical coverage is wildly diverse and uneven. Geographical and linguistic biases are strong, so you need to take this into account. Also, sometimes, when you consult open data, they unexpectedly provide inconsistent or irregular results, without any plausible explanation. We develop strategies to try and solve this, but it still is a big hindrance. There is no quality control regarding the contents Ñ it is completely unfeasible, at least in a near future. This was a surprise to me, but two years ago some people from Germany discovered bad practices regarding manipulations in Google Scholar... I don't want to give you ideas, but GS can also be manipulated, so you need to remain cautious.

In order to understand what the new generation of web indicators consists of, the best way is to establish a comparison with the key indicators in the bibliometric world.

You have indicators of activity: in bibliometrics you'll find publications, authors, institutions, in webometrics you'll have webpages, all kinds of documents (rich files in PDF format and so on), media files and profiles... Concerning impact, I mentioned before that you have citations on the one hand and links on the other. You don't only have the total number of citations, for instance you have this development by the group SCImago, who issued a Journal Ranking (SJR) that is an application of PageRank to

journals, since not all citations have the same weight. On Google, the algorithm for ranking is based on PageRank, which means that you are counting the total number of links but not every link has the same value.

For building networks, you can acknowledge and use co-authorship, co-citations, bibliographic coupling and co-words (bibliometrics tools) as well as co-inlinks and co-outlinks, mentions, coined interlinks etc. This is interesting because usage measurement is only starting to develop. The main reason is that until now there are only very few studies, for instance, about the circulation ofjournals. It is difficult to obtain data, and also to understand the impact of a journal having 10,000 copies distributed or having one million readers. The situation could change considerably in the case of web, because there you have real access to those data (visitors, visits, downloads...).

Of course you can analyze the sources. I talked about the strong biases on Google and other search engines, but during the recent years there has also been a lot of papers debating the Anglo-Saxon biases of WoS or Scopus, so...

You can develop composite indicators. You know the most important rankings of universities are based on bibliometric plus survey data. We are also using the Web to develop rankings.

This is an example. We tried to obtain the number of documents available from the OST and found more than four hundred PDF files. Perhaps it is too many, perhaps it is not enough, but the data can be obtained very easily. You can see, according to Google (that is used by 99% of the people using the Internet all over the world), that you can access more than 400 documents coming from the OST.

Regarding link analysis, we have several possibilities. You can evaluate visibility by counting the links, by counting the impact one scientist has on another, etc. via backlinks and inlinks (referring webdomains), outlinks (referred webdomains); Perhaps, in very small scenarios, we can also explore link motivations (more diverse than citation motivations) through anchor analysis , as well as link rot.

You can build networks. It is possible because we have access to this kind of information. Here is a very important point. Until 2011, we had a free access to this information provided by several of the big search engines, especially Yahoo. Unfortunately, in 2010, Yahoo and Microsoft reached an agreement, according to which Yahoo doesn’t exist as an independent database anymore and is using the data from

Bing. Therefore the system that provided links is discontinued. Now we have only a commercial system that asks money for this information. Fortunately, the quality of the new system is far better than the services provided by Yahoo (except then it was free), and the fee is not that expensive. We have agreements with several of them, and even in the current funding situation we have in Spain, we can still pay for it.

Finally, you can use your own data sources using tools you can develop by yourself (crawlers like SocSciBot for instance). This is a common strategy applied among people working in webometrics, but it isn't very popular because you need a lot of computer resources, sometimes you need to customize the system and you need to have a strong knowledge in informatics... It is not easy to use.

I'm going to introduce you to the three main providers of link data. One is American, another is English, and the third one comes from Ukraine, a very interesting new actor in this area. I'm going to focus on the results and not the providers.

Here the American system provides the list of institutions linking to OST and shows very interesting developments. You are referenced by the Observatoire des Observatoires, by the American Patent database, by the Ministry of Higher Education and Research, and of course by the European Union. Your profile looks interesting at first sight. Very important point: you don't need a subscription to access that data. You only need to register Ñ registration is easy and free Ñ to access a very basic record, but sometimes it can be enough. This slide was obtained this way. As you can see, you get very good information without spending a cent.

This is part of the most problematic part of my presentation. I'm going to try and explain why the French universities are performing well in our ranking (sic). We used the same source, and you can see I compared Harvard, Cambridge, Sao Paolo University, and UNAM in Mexico, with the Marie Curie University in Paris. There are only five because this is the free part of the systemÑ you can compare until five different systems for free. It shows that the total number of links is 16 million in the case of Harvard, five million for Cambridge, half a million for USP, almost one million for UNAM, and not even 100,000 for Paris VI. Well... This is not my data, but the data they are collecting, and the data we are using to build our ranking.

This is the second one, the British Majestic SEO site explorer. The system is very similar. Here again we researched OST visibility, and the page shows more than one thousand institutions are linking towards OST with a minimum of 10 links per institution,

which is not bad. If you think about the services you are offering, you can think: OK, there are one or two institutional links ("We are linking you because you are OST") and perhaps there are seven or eight additional links to individual reports, individual departments, or individuals, or this event, for example. Then you can think it is a rich environment, taking into account the amount and the characteristics of information the OST is offering on its website. You could think it is a low number, but it is not the case at all, except maybe in the ratio. This number can be improved.

Here we have the same situation: we have a longer list of institutions linking you. It is interesting because there are other actors involved. Not only ministries, patent databases or the European Union, but also several other institutions, for instance a Catalan one! People working in my country are linking you. Again, this report is free if you register Ñ I think the first 250 reports are free.

Part of the explanation of the problem is visible here. This is Ahrefs, the Ukrainian system. It shows the distribution of links according to the top-level referring TLDs. American Universities are not linking the EU. That's interesting too. But you have an impact on the European Union, of course in France, in Canada, Belgium (language), also in Italy and Denmark, Poland, Colombia, and again on our friends in Catalonia.

This is the kind of things we are trying to develop (CybermetricsLab, January 2012). We made an analysis of 2100 university webdomains in the world and showed the interlinking between them. As you can see there is a strong core, mainly Anglo-Saxon, since it is dominated by the U.S., the U.K., and, surprisingly, Canada. I'm sorry to say Canada is part of the American academic world... Different circles are obvious on that graph, with various small institutions, public or private, coming from South-East Asia, Latin America... But the core is North America and the United Kingdom. We are going to polish this in the future because we are trying to produce more meaningful slides... This one is quite messy.

One application regarding the use of composite indicators is the European Ranking Web of Universities... There are no French universities at all in that list...

Let's talk about altmetrics. Altmetrics were introduced by Jason Priem, a very young American researcher at the University of North Carolina. I blame him personally for coining the name of that discipline ("alt" for "alternative" doesn't mean anything here), but it has been a great success because everybody knows that today informal communication on the Web is increasingly important. We are talking about mention in

the m ajor search engines (Google, Bing, Teoma, but also, very important on a demographic point of view, the Russian Yandex, and the Chinese Baidu), in selected "closed" as groups such as newspapers, portals, Government sites, Political parties, Hospitals, Companies, etc., well as in repositories and other e-libraries.

Altmetrics are also, and especially, taking into account mentions in the Web 2.0. We are considering mentions in Wikis (Wikipedia / Wikimetrics), blogs (Blogometrics: Technoratii, Digg, ScienceSeeker), social sites

(Facebook, Google+, MySpace, Academia.edu, LinkedIn, ResearchGate), full text collections (Google Books, Amazon), multimedia (YouTube, Flickr) and Twitter (Klout, Peerindex, Twitteralyzer).

The most relevant for academic purposes is social bookmarking (Delicious, CiteULike, Mendeley, CrossRef, CitedIn, Connotea, Bibsonomy), Social Peer-review (F1000) and Social sharing (Slideshare).

The tool for making altmetrics is mention analysis. We can use traditional bibliographic units, like the author's name, with problems related to the lack of standardization (but there are proposals like ResearherID, ORCID or AuthorClaim), article (/ book/journal) title(s) Ñ which usually requires a minimum length (~35 characters) to avoid noise, institution or organizations name, etc. But there are also new units to use, for example codes (URL, DOI, Handle), when available. Today, most of the editors are provided individual DOI for each paper they publish, so it is becoming more and more popular. It is interesting because it is a unique number, therefore there is no ambiguity.

You can also use many other units. It is the case of "hot topics". They have an advantage because every hot topic has a clearly identified name that doesn't change. That way we can reduce the noise of the results. For many of the most important academic and scientific topics today, there is a key-sentence we can use (for instance "climate change"): standardized names or acronyms of supersites, portals, directories, repositories, etc. They also offer full collections or individual items such as software, slides, learning objects, or data files; mention of events like conferences, workshops, or seminars, even of email addresses (most of the scholars are using the email addresses of their institution rather that "anonymous" ones) and postal domains, or coded objects like biological scientific names, chemistry formulas, or search through

physics/astronomy datasets, because they are described in a very normalized and standardized way.

Here I'm exposing the biggest problem and limitations. Searching mentions of my own university, my Alma Mater has six million mentions, but in the Spanish official version; if you take the English version, you get about one tenth more, and if you combine both, surprisingly, the number is not a logical sum and is even larger. There are many other options you have to take into account here. The explanation for this is very simple. Google declares they support Boolean operators, but it is not true. This is part of the discoveries we made. You need to study the sources to discover their assertions are false.

Here is an example of the Mendeley social bookmarking system. In this case mention analysis can use not only the traditional mentions, but also different kinds of recognition, in this case the number of readings for Swiss-Italian researcher Benedetto Lepori. This database has 17 publications, and an h index of 7, which is pretty high since the total number of papers is only 17. You can say the numbers are still low. Yes, but this is because it is a very young product, and the numbers are growing exponentially.

Beyond mentions, you can also look at rat ings, reviews, comments, likes and shares (common features of Facebook, for instance), bookmarks, and clusters (also searched..., also visited..., friends...). Here you have some examples. This was made by British scientists studying the four top British universities, in this case using the "likes" on Facebook and the Twitter data. You can see every university has at least one profile on Facebook or one on Twitter, and you can see the numbers. As you would expect, Cambridge is in the top institutions on Facebook, since it is one of the best universities in Europe, but on Twitter Cambridge's numbers are not that different from the rest. In some cases, its impact is even below the impact of other institutions. This is an interesting point for explanation. The global impact of Cambridge is larger, because of its reputation and outputs, but for individual events Ñ this is what you can tackle with Twitter Ñ there is no reason why Cambridge should be first every time. This is a great advantage.

Jason Priem, the young scientist who coined the term of altmetrics, developed ImpactStory, a free database. You need to provide your list of publications and for each publication Ñ it can be a full paper or a collection of slides, or something like that Ñ they give a lot of indicators. The indicators they are using are distributed in two

sets: one is mostly impact on scholars, but it is also showing the impact on general public. Each of these sources refers to a different public. Here, for instance, you can see this paper has been used on Wikipedia. It's an interesting information, because you may have received several citations, that perhaps had an impact on various small groups, among colleagues working with you, but when you are also having an impact on the general public, for instance through Wikipedia, you can assume your contribution becomes more important. Of course you need to go into detail and to precise your analysis, but this is a good start, and completely free.

A more established proposal is PLoSone, already the largest journal in the world. In 2011 they published a total of six thousand papers. This year they already published ten thousand, and one of the big contributions of this system is to automatically provide article-level metrics (downloads, traditional citations, and altmetrics information). In fact it is an API that you can customize; if you are an editor, adding this interface is really simple, to obtain the same level of reach in statistics environment.

I will conclude with usagemetrics. Unfortunately, this is the newest development and it is hard to obtain data to analyze, but I think it is very promising. There are various ways to obtain data. You can update data directly from the server (which can be a problem because of privacy issues) if you have access to the log-files. These logs are huge Ñ during a month, you can get millions of visitors Ñ and these are files that require specialized analytical software. You can use external tools, the most important being Google Analytics. It is free and provides a lot of information, but its main advantage is that it is becoming the standard. If you are publishing your data, or a set of your data, obtained from Google Analytics, other people using this system can compare it, because the system is working the same way. Is Google Analytics the best system? Probably. I don't know, but it is the *de facto* standard, because most people are using it. Another way of getting information is being a registered user of specially developed in-house or commercial software, that are mostly private. The American MESUR project is using data from several universities, contracting these electronic libraries to several providers and collecting a lot of information. That is another source you can use.

Usagemetric is the discipline, usage analysis is the tool. You can analyze the traffic, the global presence of your institution. There are several options but only one source is really valid and secure, although not totally reliable; it is called *Alexa* and it is strongly biased from a geographical point of view, but it is still possible to make an analysis. You can also count data directly from your profile (number of visitors, number of visits, origin

of the visitors, world and national rank according to the mean number of visits received in the last three months by the whole webdomain, etc.). Perhaps the most important is that you can see the terms and codes the visitors used to find you, which is a very rich environment, from a semantic perspective, but also from a practical point of view. You can see not only the top words they are using to reach you, but especially the trail that lead them to you.

Here are two samples. This is Alexa's profile for Paris VI University. Links is a far better indicator than usage, but usage is also very interesting: again, the number of visits is far below the four other competitors', and here is the reason why if you look at the geographical impact of its outputs through Regional Traffic Ranks: France, Tunisia, Algeria, Morocco... Most of the information is provided in French. Perhaps this is part of the problem. There is not enough international information coming out of French universities. And you know I am referring to English.

This is Google Analytics' page for Paris VI. Let me remind you that Google Analytics is providing 200 different indicators. Not all of them are useful for us, but they offer a tremendous amount of information.

We are going to change the moto of research from "Publish or perish" to "Web-publish or perish". Why not? There is a lot of research and publishing on the Web. If you do that, clearly, the quality of the tools I am introducing here are going to better reflect the true situation. Here I am proposing a combination of link, mention and usage analysis, used in conjunction with citation analysis to extend (large scale populations), accelerate (immediacy) and deepen (increased diversity of actors and actions) the description of scholarly communication processes.

Problems remain, of course. We are working on the reliability of the data, of the data sources, and apart from the theoretical developments, more research is required about the data sources, specially regarding the quality of the information (sometimes it is unreliable, biased and frequently noisy) and the way the info is obtained, indexed and provided (most of the procedures are opaque). But in the end, and this is the main reason I am here, I want to show you that there are enormous opportunities for working in this area, because we are offering a set of tools that are answering traditional problems that weren't that easy to address using bibliometrical tools.

That is all, thank you very much for your attention.