

Web Services for Bibliometrics

Sylvie Godel, Isabelle de Kaenel, Pablo Iriarte

Medical Library, University Hospital Center – Lausanne, Switzerland

sylvie.godel@chuv.ch

Abstract

Institutional repositories have spread in universities where they provide services for recording, distributing, and preserving the institution's intellectual output.

When the Lausanne “academic server”, named SERVVAL, was launched at the end of 2008, the Faculty of Biology and Medicine addressed from the outset the issue of quality of metadata. Accuracy is fundamental since research funds are allocated on the basis of the statistics and indicators provided by the repository. The Head of faculty also charged the medical library to explore different ways to measure and assess the research output.

The first step for the Lausanne university medical library was to implement the PubMed and the Web of Science web services to easily extract clean bibliographic information from the databases directly into the repository.

Now the medical library is testing other web services (from CrossRef, Web of Science, etc.) to generate quantitative data on research impact mainly. The approach is essentially based on citation linking.

Although the utility of citation and bibliometric evaluation is still debated, the most prevalent output measures used for research evaluation are still those based on citation analysis. Even when a new scientific evaluation indicator is proposed, such as h-index, we can always see its link with citation. Additionally, the results of a new indicator are often compared with citation analysis. The presentation will review the web services which might be used in institutional repositories to collect and aggregate citation information for the researchers' publications.

Introduction

When the Lausanne institutional repository, named SERVAL (<http://serval.unil.ch>), was planned during the year 2007, the Faculty of Biology and Medicine (FBM) emphasized from the outset the importance of metadata quality. To face this legitimate concern, the repository project team, which included representatives from the university medical library, resolved to rely as much as possible on standardized information and authority files. It was principally decided to integrate and exploit the unique identifiers generated by scientific publishers and databases producers. Three main providers were analyzed: CrossRef for the Digital Object Identifier (DOI), National Library of Medicine for the PubMed Identifier (PMID) and Thomson Reuters for the Web of Science (WoS) Unique identifier (UT). It was assumed that these numbers and codes, associated with automated services, would facilitate the regular transfer and update of external reliable metadata into the local repository.

Besides, the quality concern expressed by the Faculty was linked to another request addressed, not to the repository project team, but directly to the medical library. The Head of Faculty mandated the library to study how to perform metric analysis using the data gathered in the repository.

The aim was to use the deposited records in order to assess the publication activity of the FBM research community, at both group and individual level. As scientific publication represents a significant part of the research process and output, this assessment is fundamental for research funds allocation, grants decisions, policy making and individual promotion.

Until now, the evaluation process performed at the FBM would usually take into account various criteria:

- **Number of publications** over time
- **Impact factors (IF)** of the journals which the researcher has published in, provided by Thomson Reuters Journal Citation Reports® (JCR). A journal's IF is the ratio between the number of current year citations and the source items published in that journal during the previous two years(1).
- **Research Production Unit (RPU)**: indicator derived from the journal IF and pondered by domain in order to increase the homogeneity of subfields. It's calculated with the formula $RPU = 10(1 - e^{(-IF/x)})$ "where IF is the impact factor of the journal and x the mean IF for the subfield in which the journal belongs"(2).
- **IF and RPU pondered with the type of publication** (letters, reviews and case reports have less weight than original articles) **and with the degree of contribution** (is the researcher ranked first? last? or in the middle?).

In 2005, Hirsh published his seminal article about the **h-index** and this new indicator received a great attention from the research community. According to Hirsch, a scientist has index h if h of his or her papers have at least h citations each and the other papers have h citations or less(3).

With the h-index, the quality of output is measured using citations counts at article level. The Lausanne FBM rapidly recognized the h-index as a simple yet sound estimator of the research output of individual. The mandate for the library was to analysis how to include the h-index calculation among other indicators by making use of the metadata stored in the repository.

Background

The SERVAL deposit workflow was planned to rely to a large extent on the commitment of end-users. The challenge was to assist the creation of records in order to help the submission process and to ensure a high level of data quality. It was intended to reduce typing errors or multiple key strokes, such as the tedious « copy/paste » combination. The repository project team started to analyze possible sources of reliable biomedical metadata that could be incorporated into the repository.

In the scientific field, clean, authoritative and accurate bibliographic datasets are available from various providers: bibliographic databases, cited reference-enhanced systems, publishers' sites, library catalogs, controlled lists and repertories. These different platforms usually facilitate data transfer and integration into local systems, through many channels and applications. Among the current technologies available are:

- **Export functions:** mainly designed to push a set of records into personal reference software that supports various standards (RIS, MEDLINE, BibTeX, etc.) and that can generate files for upload into local applications such as repositories.
- **Open URLs and Link resolvers** which allow pushing a single record into an entry form on a target server.
- **Web services** technology that appeared more recently and that can be used for a lot of applications: single record creation and completion, batch input routines, mashups, etc.

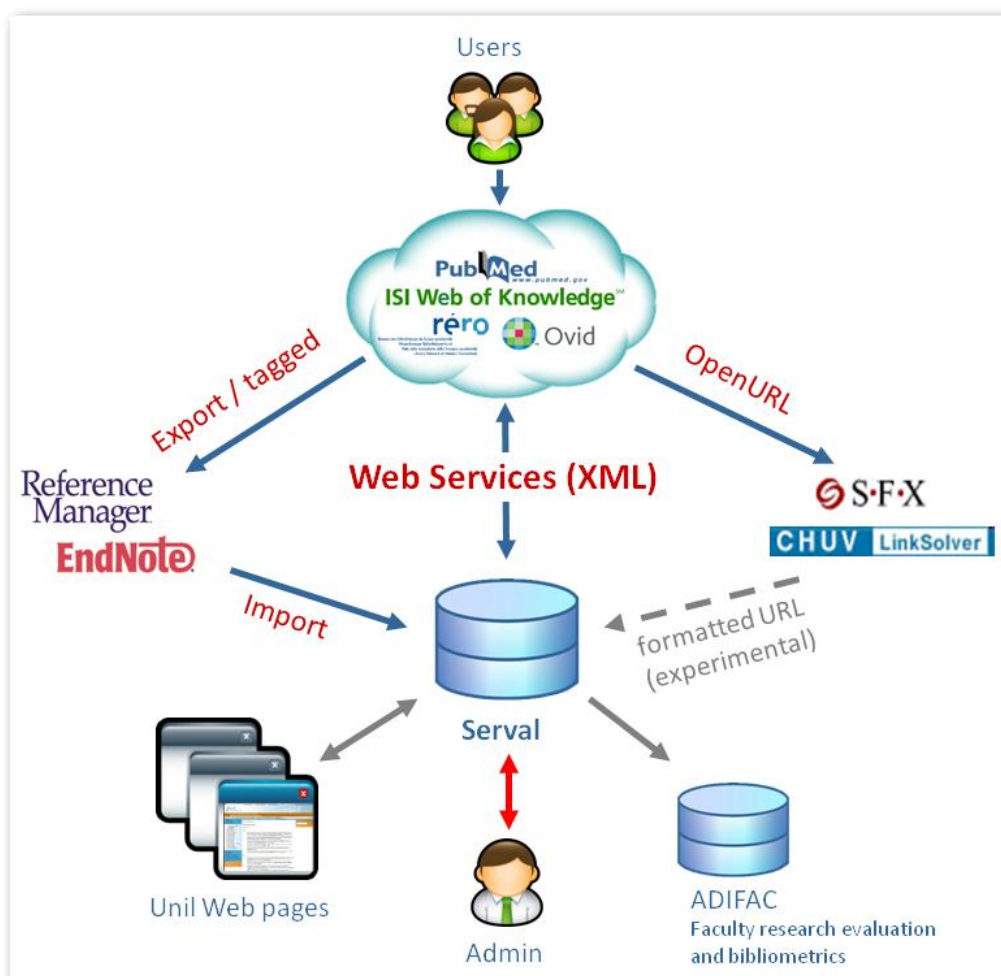


Fig. 1 The metadata acquisition techniques tested in SERVAL

These technical features were not equally supplied by all providers. Very early on, during the study process, the technical features offered by PubMed were analyzed. PubMed, produced by the National Library of Medicine (NLM) in the United States, represents one of the largest biomedical bibliographic databases in the public domain. It presently indexes more than four thousand international medical journals. For each publication, the database gives access to a wide range of raw data: complete list of authors, affiliation for the first author, abstract, publication type, ISSNs, DOI, etc. In 2002, NLM enhanced the functionalities of the Entrez search engine and the Application Program Interface (API) called E-Utilities. Combined with the AJAX (Asynchronous JavaScript and XML) technology, those web services turn out to be particularly efficient and can be implemented in the user environment or front-ends in order to assist the researchers at the metadata point of entry, filling in automatically the bibliographic metadata fields on the repository web form. In consequence, it was easy to choose PubMed as one the main source of metadata for SERVAL.

Unfortunately, the NLM model in facilitating a direct transfer of bibliographic information into local servers was not immediately followed by the other major commercial providers of STM resources, such as Thomson Reuters and Elsevier. In 2004, the citation-enhanced database SCOPUS was launched by Elsevier with some embryonic web service. In 2009 only, the year SERVAL was launched, Thomson Reuters changed attitude, and developed web services and far-reaching online tools facilitating the access of the Web of Science (WoS) content, allowing their data to be easily integrated into a custom application such as a repository.

For those journals not incorporated in PubMed, WoS or SCOPUS, the CrossRef database also offers a web service that can be used by the libraries without cost.

Taking all these characteristics into consideration, the SERVAL project team took three steps to assist the metadata creation of biomedical references in the repository so as to ensure a high level of quality and the inclusion of a maximum of identifiers, guarantee for an optimal re-use of the metadata for present and future needs like research assessment or bibliometrics.

Promoting the use of web services to automatically populate the repository

When connected to a personal account, the researcher calls the record entry form. Then he can fill it, just by typing in the PubMed, Web of Science, or CrossRef unique identifier, respectively called PMID, UT and DOI. This technique is recommended during training and tuition of end-users.

In the background, the system makes an AJAX callback to the web service corresponding to the identifier, parses the XML response sent by the provider and then it filters and maps metadata fields prior to introduction into the repository(4).

Of course, this solution means that the researcher has looked up for the relevant identifiers before he starts capturing the external content and pulls the records directly into a personal account. Concerning single-item deposit, the alternative is to use the link resolver, which is implemented in the major databases and can “push” the main fields of the references directly into the repository web form. But in many cases, the link resolver solution is poorer in terms of metadata that can be transferred through the OpenURL format. As a matter of fact, the OpenURL specification is intended to identify the resource and not to carry the secondary data like abstract, keywords, complete list of authors, authors address, etc. Finally, manual entry is used only for those publications that cannot be retrieved from international abstracting and indexing (A&I) databases.

Implementing alerts and batch imports

It was also decided that repository staff will carry out deposit on behalf of the departments and authors. Under the medical library supervision, regular batch imports of bibliographic records from authoritative external biomedical databases are performed. Publications of the FBM research community are identified through alerts placed on PubMed and Web of Science. The search terms include variations of affiliation denominations. WoS yields more records as the database includes the affiliations for all the authors. Every week, batches of records are entered by the repository content managers after a rapid analysis of the coherence of the retrieved set. Of course, the inconsistencies of the affiliation designations submitted by the authors to the journals may affect the publications recognition. Indeed, the bibliographic systems derive the institution addresses from information harvested from the publishers' sites. The FBM research office regularly gives instructions and recommendations concerning a standardization of the affiliation, but the local researchers are not very respectful of these guidelines(5).

Each record entered by the content managers at the institution level has to be attributed to one (or many) research unit(s), and to one (or many) faculty researcher(s), depending on the internal collaborations that took place for the production of the paper. This process is assisted by a standardized proposition list, but author disambiguation needs a human intervention so as to attribute the publication to the right person(s) working in the right service(s).

A collaborative approach to metadata control

To sum up, the SERVAL policy promotes a mediated deposit: the authors and the library staff can upload or enter records, but there is a validation on both sides. Academics have to approve (or decline) lists of references attributed to them after a batch import. A library staff member, appointed as repository administrator, has to validate the metadata deposited by the authors. Among control tools, SERVAL offers an automatic detection of duplicates, audit trails of changes, and reporting applications to keep track of the modifications for each record.

This collaboration between managers, administrator and end-users has to be patiently built through training sessions, meetings and guidelines and is not always as smooth as could be expected. It is not always clear for the researchers, why they should ensure that correct metadata are stored for them in the repository.

Objective

As a result of the workflow described above, the collected records generate fairly standardized and controlled lists of publications for the different members of the research community. The logic step forward is to analyze if these consolidated data available in the repository would provide an adequate basis to perform quantitative measures and bibliometrics. The faculty research department mandated the medical library to explore the new ways that can improve the current methods of bibliometric analysis and to add citations counts into the metadata used to compare the faculty staff. The current assessment process is divided into two parts:

1. **General indicators of the institutional research output:** The publications lists are harvested from SERVAL and added to the faculty administration management tool (ADIFAC). This tool can merge the publications data with the journals IF information and calculate the research trends across the time for the whole faculty, down to the research unit level.
2. **Personal indicators at the individual level:** bibliometric reports are produced both for internal promotion and for evaluation of external candidates' applications. The bibliometric analysis is run by the faculty research evaluation unit, with the help of the medical library. Only the publications of the last 5years are evaluated. As already mentioned in the introduction, the evaluation process takes into account various criteria (IF, RPU, etc.). The results are then used to benchmark the faculty staff or external candidates, comparing them with the average data of people in the same domain (clinicians, fundamental researchers or psychiatrists) and with the same professional level.

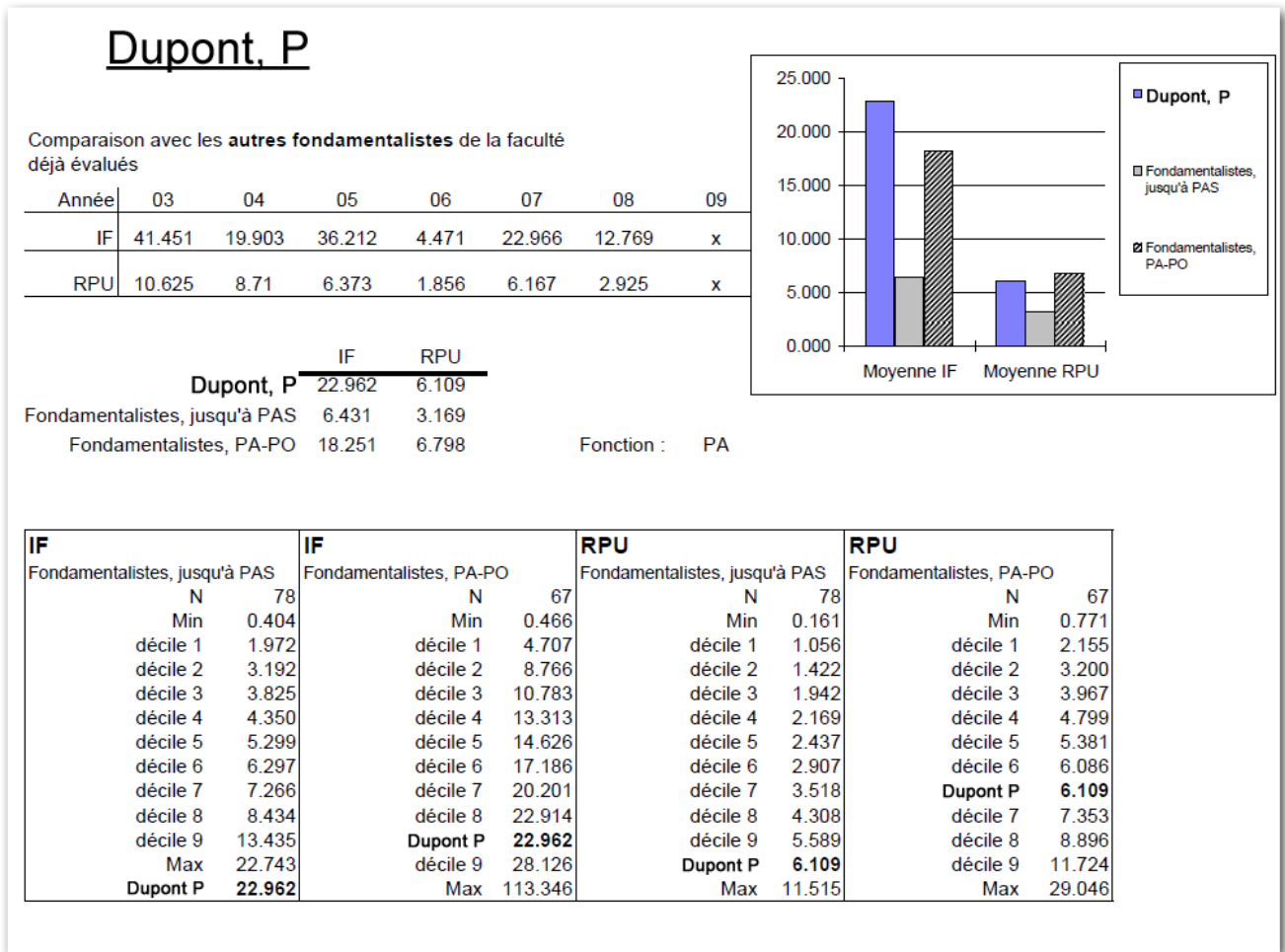


Fig. 2 A bibliometric benchmark at the FBM

At the moment, the two kinds of assessment are using only one kind of external "scale value", even though it's weighted and normalized by some mechanisms. This scale, derived from the Impact Factor (IF), measures only the importance of a selected group of journals. There are a lot of journals without IF or in the pipeline, expecting to enter into the club like many Open Access journals.

There are two alternatives to the IF: the SCImago Journal Rank (SJR) indicator(6) and the Source Normalized Impact per Paper (SNIP)(7).

Based in the Google PageRank algorithm that takes care of the structure of the citations maps (some citations are more valuable than others), the SCImago Journal Rank (SJR) measures the visibility of

the journals contained in the SCOPUS database but it takes into account only the articles published after 1996. The SJR is supported by Elsevier and their database SCOPUS, the principal competitor of Thomson Reuters Web of Science. In contrary to the JCR, the SCImago database is entirely free and people can download the complete list(6).

Created at CTWS, University of Leiden, by Professor Henk Moed, the Source Normalized Impact per Paper (SNIP) "measures contextual citation impact by weighting citations based on the total number of citations in a subject field. The impact of a single citation is given higher value in subject areas where citations are less likely, and vice versa"(8).

Some statistical criticisms persist concerning the three systems, IF, SJR and SNIP, since they all give an average or probabilistic hope of citations for all the papers of a journal. In fact, some studies have shown that in most cases 20% of papers takes the 80% of citations(9). In consequence, a high impact or prestigious journal can be the distorted result of many citations of a few papers rather than the average level of the majority. In this respect, the IF has a limited value as an objective measure of individual papers(10). The measurement of research performance at the level of the individual scientist remains problematic and requires a new kind of metrics.

Among these new metrics, the h-index proposed in 2005 by Hirsch(3), drew a great deal of interest within the research community. In 2006, Egghe introduces the g-index, an improvement of the h-index taking into account the global citation performance of a set of articles: "g is the largest rank (where papers are arranged in decreasing order of the number of citations they received) such that the first g papers have (together) at least g^2 citations"(11). Both indicators requires a citedness score for each individual record and could only be derived from a large cited reference-enhanced database like Web of Science, SCOPUS or Google Scholar. They are very difficult to integrate it in a management tool or database because they are expected to move, the citation counts changes over time!

Method

The experience acquired implementing bibliographic metadata acquisition in the repository has helped to follow the path, now seeking to integrate bibliometric information with the repository metadata via the web services. This method allows calculating the number of citations per publication, the h-index and the g-index of a researcher of the faculty on the fly.

The first step was to identify and compare the different bibliographic resources containing citation information that can be consumed through a web services protocol.

Web of Science (<http://www.isiknowledge.com>)

Web of Science (WoS), the Thomson Reuters citation enhanced database is historically the first, and still the largest bibliographic database of its kind with 12'000 journals, 46 millions of master records and more than 750 millions of cited references from 1900 till now(12). Institute for Scientific Information, now called Thomson Reuters, pioneered citation analysis tools with its databases and the calculation of the impact factor for the journals. It still offers information that complements other databases, in terms of journal coverage and information capture. Among the major characteristics of WoS are, on the one hand, a multi-valued affiliation field (in comparison to PubMed which only signals the first author affiliation) and, on the other hand, a citation count for each record.

Three web services have been launched in 2009 by Thomson Reuters: the "ISI Web of Knowledge Web Services"(13), the "Article Match Retrieval Service (AMR)"(14), and an OpenURL resolver. They are all earmarked for subscriber institutions (IP authentication). In Switzerland all universities have been subscribing for many years.

The "ISI Web of Knowledge Web Services" is the complete version, it returns rich metadata in XML to different queries but it's quite complex to configure and needs SOAP advanced skills. On the opposite, the light version LAMR proved to be good choice for the purpose sought in Lausanne: simple to implement (only a HTTP POST form to configure), flexible, including advanced query options. Besides, it returns the essential information needed to calculate the h-index(15):

- Citation counts
- Web of science unique identifier "UT". This identifier can be used in combination with the complete version of the web service to retrieve the rest of the metadata.
- URLs to make deep links to the master record and to the "cited by" references into Web of science

```

Response :
<?xml version="1.0" encoding="UTF-8" ?>
<response xmlns="http://www.isinet.com/xrpc41" src="app.id=SERVAL_Web_Services,env.id=Bibliometrics,partner.email=pablo.iriarte@chuv.ch">
<fn name="LinksAMR.retrieve" rc="OK">
<map>
<map name="cite_id">
<map name="WOS">
<val name="relatedRecordsURL">
<![CDATA[http://gateway.isiknowledge.com/gateway/Gateway.cgi?GWVersion=2&SrcApp=PARTNER_APP&SrcAuth=LinksAMR&KeyUT=A1984TS61800024&
DestLinkType=RelatedRecords&DestApp=ALL_WOS&UsrCustomerID=2109fe80815e8621515050a12879ac1a]]></val>
<val name="sourceURL">
<![CDATA[http://gateway.isiknowledge.com/gateway/Gateway.cgi?GWVersion=2&SrcApp=PARTNER_APP&SrcAuth=LinksAMR&KeyUT=A1984TS61800024&
DestLinkType=FullRecord&DestApp=ALL_WOS&UsrCustomerID=2109fe80815e8621515050a12879ac1a]]></val>
<val name="timesCited">897</val>
<val name="ut">A1984TS61800024</val>
<val
name="citingArticlesURL"><![CDATA[http://gateway.isiknowledge.com/gateway/Gateway.cgi?GWVersion=2&SrcApp=PARTNER_APP&SrcAuth=LinksAMR&K
eyUT=A1984TS61800024&DestLinkType=CitingArticles&DestApp=ALL_WOS&UsrCustomerID=2109fe80815e8621515050a12879ac1a]]></val>
</map>
</map>
</map>
</fn>
</response>

```

Fig. 3 The WoS Article Match Retrieval Web Service response

These new retrieval functions imply a major breakthrough for those wanting to enrich the metadata of local repositories and to perform bibliometric evaluation for an institution. Thomson still retains a competitive advantage over the new competitors: Elsevier with SCOPUS and Google with Scholar for example(16). However, the implementation of the Web of Science API reveals that some metadata are not easy to capture. For example the ISSN is always missing in the complete XML file. Besides, the list of publication types implemented in the system is not suitable for the Lausanne FBM evaluation needs. Different databases imply different granularity and types of metadata. For example, the two systems (PubMed and WoS) use different document type categories with only a few in common (article, review, and letter are the most important types that are shared).

In 2004, Elsevier launched SCOPUS, a huge database with bibliometric functionalities able to compete with Web of Science. SCOPUS index the content of 18'000 titles (including more than 1'200 Open Access journals), 350 book series and 3.6 millions of conference papers, so near to 40 million records at the moment(17). The information about citations concerns only the records after 1996 (20 million, 78% with references). The other half (20 million of pre-1996 records) was captured without references and go back as far as 1823.

Very early on, SCOPUS offered a web service with a free version of their API(18) including citation counts. Some metadata elements like the abstract or the complete list of authors are available only for the subscribing institutions (the University of Lausanne does not subscribe at the moment). The free version of the API requires a key in order to authenticate all the queries. This API key can be obtained registering onto the SCOPUS web site. Each key can be used only from a single web site but 5 different keys are delivered at no cost.

After 5 years, this web service has not developed significantly and the documentation of the API is not substantial and detailed. For example, the information concerning the different formats of the response (XML or JSON) is very poor and the examples can be found only in external blogs(19). The most important limitation for people interested in this free version of the web service is that the non subscriber institutions can only make 10 requests by minute. However, the query possibilities are very large and the web service returns also the citation counts allowing the calculation of the h-index(20):

- Citation counts
- URL to make a deep link to the reference into SCOPUS

```
Request:
http://www.scopus.com/scsearchapi/search.url?devId=[the developer ID]&search=PMID%286091913%29&callback=t

Response:
t({"PartOK":{"TotalResults":"1","ReturnedResults":"1","Position":"0","Results":[
  {"title":"Organization of the higher-order chromatin loop: Specific DNA attachment sites on nuclear scaffold",
   "firstauth":"Mirkovitch, J.",
   "citedbycount":"391",
   "pubdate":"1984",
   "sourcetitle":"Cell",
   "inwardurl":"http://www.scopus.com/inward/record.url?eid=2-s2.0-0021675784&partnerID=65&md5=4b9314f61428df989c6b434005a7dc36
  }
]})
```

Fig. 4 The SCOPUS Web service request and response

PubMed (<http://www.pubmed.org>) and PubMed Central (<http://www.pubmedcentral.nih.gov>)

NLM was a pioneer in many fields and the web services development is no exception. NLM was one of the first database producers offering a set of innovative public APIs, the "Entrez Programming Utilities"(21) which are different tools providing automated retrieval options for Entrez databases data, free for all and very well documented. There are different PubMed and PubMed Central Web services that could be used for a lot of projects. For example, the "ESearch" web service can run a complex query and then retrieve the PubMed Identifiers (PMIDs) of the documents returned. Then we can

operate the “EFetch” or “ESummary” utilities to obtain all the metadata in a machine readable format like XML.

The creation of PubMed Central Open archive in February 2000 brings a new dimension to the metadata stored in the PubMed database. In fact, the bibliographies of papers deposited in the archive began to be accessible and searchable. Now, many of the two millions of articles stored in PMC gives the kind of information that can be exploited by the bibliometrics methods. For example, the EFetch Web service can be use with PubMed Central identifier (PMCID) and then it delivers the metadata and the full text of the article in a XML format(22). At the moment there is no web service in PMC returning explicitly the citations counts for a given identifier like WoS or SCOPUS but this number can be obtained by parsing the XML results of a query that returns all the PMIDs of papers citing a given document:

```
Request:
http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed&cmd=neighbor&linkname=pubmed_pmc_refs&id=6091913

Response:
<?xml version="1.0"?>
<!DOCTYPE eLinkResult PUBLIC "-//NLM//DTD eLinkResult, 10 August 2009//EN" "http://www.ncbi.nlm.nih.gov/entrez/query/DTD/eLink_090910.dtd">
<eLinkResult>
  <LinkSet>
    <DbFrom>pubmed</DbFrom>
    <IdList>
      <Id>6091913</Id>
    </IdList>
    <LinkSetDb>
      <DbTo>pmc</DbTo>
      <LinkName>pubmed_pmc_refs</LinkName>
      <Link>
        <Id>2774341</Id>
      </Link>
      <Link>
        <Id>2683591</Id>
      </Link>
      [...]
    </LinkSetDb>
  </LinkSet>
</eLinkResult>
```

Fig. 5 The PubMed Web service request and response with the identifiers of Pubmed Central articles citing a given PMID

Citebase (<http://www.citebase.org>)

"Citebase Search is a semi-autonomous citation index for the free, online research literature. It harvests pre- and post- prints (most authors self-archived) from OAI-PMH compliant archives, parses and links their references and indexes the metadata in a search engine"(23).

This database covers mostly physics, mathematics, information science, and biomedical papers published by BioMed Central or archived in PubMed Central. It was associated with ArXiv in order to provide metrics for citations, links and downloads of the ArXiv material.

Like PubMed Central, the data are accessible through an OAI interface (<http://citebase.eprints.org/cgi-bin/oai2>) and if the identifier of a publication is known (it could be the PubMed Central or BioMed Central identifier), then the complete metadata are available in several XML formats including the list of identifiers for the papers citing the publication.

Request:

```
http://www.citebase.org/oai?verb=GetRecord;metadataPrefix=amf;identifier=oai:pubmedcentral.gov:1379635
```

Reponse :

```
<?xml version='1.0' encoding='UTF-8'?>?xml:stylesheet type='text/xsl' href='http://www.citebase.org/static/stylesheets/celestial.xsl'?>
<OAI-PMH xmlns='http://www.openarchives.org/OAI/2.0/' xmlns:xsi='http://www.w3.org/2001/XMLSchema-instance'
xsi:schemaLocation='http://www.openarchives.org/OAI/2.0/ http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd'
<responseDate>2010-06-10T00:20:56Z</responseDate>
<request identifier='oai:pubmedcentral.gov:1379635' verb='GetRecord' metadataPrefix='amf'>http://www.citebase.org/oai</request>
<GetRecord><record><header><identifier>oai:pubmedcentral.gov:1379635</identifier><datestamp>2006-02-25</datestamp></header>
<metadata><amf xmlns='http://amf.openlib.org' xmlns:xsi='http://www.w3.org/2001/XMLSchema-instance' xsi:schemaLocation='http://amf.openlib.org
http://amf.openlib.org/2001/amf.xsd'><text id='oai:pubmedcentral.gov:1379635'>
<title>Prevalence of cardiovascular risk factors in a middle-income country and estimated cost of a treatment strategy</title>
<displaypage>http://www.pubmedcentral.gov/articlerender.fcgi?artid=1379635</displaypage>
<copyright>Copyright © 2006 Bovet et al; licensee BioMed Central Ltd.</copyright>
<hasauthor><person><name>Bovet, Pascal</name></person></hasauthor>
<hasauthor><person><name>Shamlaye, Conrad</name></person></hasauthor>
<hasauthor><person><name>Gabriel, Anne</name></person></hasauthor>
<hasauthor><person><name>Riesen, Walter</name></person></hasauthor>
<hasauthor><person><name>Paccaud, Fred</name></person></hasauthor>
<serial><endpage>9</endpage><issuedate>2006</issuedate>
<journaltitle>BMC Public Health</journaltitle><startpage>9</startpage><volume>6</volume></serial>
<reference><literal>Ezzati, M; Lopez, AD; Rodgers, A; Vander Hoorn, S; Murray, CJL (2002) "Selected major risk factors and global and regional burden of
disease" Lancet 360 1347-1360
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&amp;db=PubMed&amp;dopt=Abstract&amp;list_uids=12423980</literal></reference>
[...]
<reference><isreferencedby><text ref='oai:pubmedcentral.gov:1501045' /></isreferencedby>
</text></amf></metadata>
</record></GetRecord></OAI-PMH>
```

Fig. 6 The Citebase Web service XML request and response with citation information

CiteSeer^x (<http://citeseerx.ist.psu.edu>)

This computer and information science digital library is also an innovative and experimental platform providing new citation analysis methods and algorithms to parse bibliographies of the PDF documents founded on the web.

Like other Open Archives it can be explored using the OAI-PMH protocol. If the internal identifier of a document is picked out (unlike Citebase, external identifiers like PMCID cannot be used), then all the metadata in an XML format can be obtained, including the identifiers of the documents citing it.

Google Scholar (<http://scholar.google.com>)

Google Scholar could represent a serious alternative, only if the sources accessed by Google become more transparent and reliable. At the moment, despite the high demand, there's no web service and any effort to filter the HTML response of Google Scholar in order to extract the citation counts are blocked by Google very quickly. It is possible to use a Firefox extension(24) or the Publish or Perish (PoP) software(25) in order to compare the h-index of Web of Science and SCOPUS with the Google Scholar h-index(26).

CrossRef (<http://www.crossref.org>)

CrossRef aims to be the "citation linking backbone for all scholarly information in electronic form". In that sense and through CrossRef Digital Object Identifiers (DOI) they have built one of the largest bibliographic metadata databases in the world (40 millions of DOIs registered at present). But CrossRef is also a "collaborative reference linking service that functions as a sort of digital switchboard. It holds no full text content, but rather effects linkages, which are tagged to article metadata supplied by the participating publishers. The end result is an efficient, scalable linking system through which a researcher can click on a reference citation in a journal and access the cited article"(27).

In 2004, CrossRef and Atypon launched "CrossRef forward linking", a service that allows publishers members of CrossRef to know if their publications are being cited and to incorporate that information directly into their online publication platforms. This service is free of charge for the publishers but, "in order to participate, there is an important quid-pro-quo: in order to discover what publications cite your content, you must in turn submit metadata listing the works that your publications cite"(28).

The metadata and the identifiers of bibliography citations can be included in the process of a DOI deposit. This means that CrossRef has entered into the market of citations counts and "who cites who". They could potentially threaten the two major reference-enhanced databases: Web of Science and SCOPUS. At the moment, this information remains confidential and only publishers' members can partially make use of them. It could be an outstanding step if this information became accessible for the academic libraries community and could be accessed via web services. Only then, CrossRef citation counts and links to citing articles could be included in the academic repositories.

There are a lot of other bibliographic databases or Open Archives including bibliometric information or citation counts: RePEc, CINHALL, PsycINFO, PROLA, etc. Unfortunately their content is too limited and the citation counts are insufficient to calculate the h-index(29).

The essential role of identifiers: ISSN, ISBN, DOI, PMID, PMCID and UT

The assessments run in Lausanne FBM are based on matching the publication metadata with the IF of the journal given by the ISI Journal of Citation Report (JCR). Unfortunately, the web service allows to query the JCR database by ISSN but returns only the URL of the deep link to the Impact Factor Trend, instead of the IF himself. Therefore the IF has to be taken out of an internal database containing the data of the last edition of the JCR CD-ROM.

The ISSNs introduced in SERVAL, usually imported from external databases like PubMed, facilitates this operation, but sometimes the ISSNs are different since each database chooses the ISSN version judged convenient. Usually, JCR takes the ISSN attached to the print version but PubMed has chosen the ISSN of the version used in the indexing process. Presently, the electronic version ISSN is retained by PubMed in the most cases. Fortunately, PubMed has recently introduced the ISSN linking (ISSN-L) as secondary ISSN for the whole database. The ISSN-linking and the ISSN print are usually the same. In order to ensure this matching process the whole ISSN table that gives all the ISSN forms for each ISSN-L entry has been downloaded from the www.issn.org portal. Then, this information was merged with the JCR data and the PubMed list of journal downloaded from http://www.nlm.nih.gov/tsd/serials/terms_cond.html. Now this table is used to match the ISSN of the SERVAL records with the IF of the journal in the bibliometric process.

ID JCR	ID PubMed	Title JCR	Title PubMed	Abbr JCR	Abbr PubMed	ISSN JCR	ISSN PubMed	E-ISSN PubMed	Impact Factor JCR
1	22696	AAPG BULLETIN	AAPG bulletin	AAPG BULL	Am Assoc Pet Geol Bull	0149-1423	0149-1423		01.273
3	30088	AAPS PHARMSCITECH	AAPS PharmSciTech	AAPS PHARMSCITECH	AAPS PharmSciTech	1530-9932		1530-9932	01.351
5	2280	ABDOMINAL IMAGING	Abdominal imaging	ABDOM IMAGING	Abdom Imaging	0942-8925	0942-8925	1432-0509	01.213
8	20159	ACADEMIC EMERGENCY MEDICINE	Academic emergency medicine : official journal of the Society for Academic Emergency	ACAD EMERG MED	Acad Emerg Med	1069-6563	1069-6563	1553-2712	01.990
9	1346	ACADEMIC MEDICINE	Academic medicine : journal of the Association of American Medical Colleges	ACAD MED	Acad Med	1040-2446	1040-2446	1938-808X	02.571
10	20370	ACADEMIC RADIOLOGY	Academic radiology	ACAD RADIOL	Acad Radiol	1076-6332	1076-6332	1878-4046	02.094
11	20263	ACCOUNTS OF CHEMICAL RESEARCH	Accounts of chemical research	ACCOUNTS CHEM RES	Acc Chem Res	0001-4842	0001-4842	1520-4898	16.214
26	32803	ACOUSTICS RESEARCH LETTERS ONLINE-ARLO	Acoustics research letters online : ARLO	ACOUST RES LETT ONL	Acoust Res Lett Online	1529-7853		1529-7853	01.083
27	33137	ACS Chemical Biology	ACS chemical biology	ACS CHEM BIOL	ACS Chem Biol	1554-8929	1554-8929	1554-8937	04.741
28	34787	ACS Nano	ACS nano	ACS NANO	ACS Nano	1936-0851	1936-0851	1936-086X	
31	31677	ACTA AGRICULTURAE SCANDINAVICA SECTION A-ANIMAL SCIENCE	Acta agriculturae Scandinavica Section A, Animal science	ACTA AGR SCAND A-AN		0 0906-4702	0906-4702	1651-1972	00.425

Fig. 7 the journals table merging the data from WoS and PubMed using the ISSN-L

The ISBN is also included in SERVAL books metadata and it's very important to retrieve additional metadata, cover images or links to the digital version automatically via the Web Services of the Library of Congress, WorldCat, Amazon, Google Books Search, etc. This identifier has a minor role in the bibliometric analysis of academic publications since evaluation essentially takes into account the journal articles. The citations of books are not easily workable. However, some databases, like Google Scholar, provide citation counts for books.

[\[LIVRE\] The structure of scientific revolutions](#)
 TS Kuhn - 1970 - svn.assembla.com
 Preface The essay that follows is the first full published report on a project originally conceived almost fifteen years ago. At that time I was a graduate student in theoretical physics already within sight of the end of my dissertation. A fortunate involvement with an experimental college ...
 Cité 41789 fois - [Autres articles](#) - [Obtenir@UNIL](#) - [Find in RERO](#) - [Les 54 versions](#) - [Importer dans RefMan](#)

Fig. 8 Citation counts of a book in Google Scholar

Cited Author	Cited Work [SHOW EXPANDED TITLES]	Year	Volume	Page	Article ID	Citing Articles **	View Record
KUHN	STRUCTURE REVOLUTION	1970				1	
KUHN	STRUCTURE SCI REVOLT	1970				1	
KUHN	STRUCTURE SCI REVOLU	1970				14	
KUHN	STRUCTURE SCI REVOLU	1970		2		1	
KUHN	STRUCTURE SCI REVOLU	1970		10		1	
KUHN	STRUCTURE SCI REVOLU	1970		21		1	
KUHN	STRUCTURE SCI REVOLU	1970		77		1	
KUHN	STRUCTURE SCI REVOLU	1970		101		1	
KUHN	STRUCTURE SCI REVOLU	1970		170		2	
KUHN	STRUCTURE SCI REVOLU	1970		175		1	
KUHN	STRUCTURE SCI REVOLU	1970		176		1	
KUHN	STRUCTURE SCI REVOLU	1970		191		1	
KUHN	STRUCTURE SCI REVOLU	1970		195		1	
KUHN	STRUCTURE SCI REVOLU	1970		204		1	
KUHN	STRUCTURE SCI REVOLU	1970		208		1	
KUHN	STRUCTURE SCI REVOLU	1970		209		1	
KUHN	STRUCTURE SCI REVOLU	1970		CH5		1	
KUHN	STRUCTURE SCIENTIFIC	1970				9	
KUHN	STRUCTURE SCIENTIFIC	1970		4		1	
KUHN	STRUCTURE SCIENTIFIC	1970		158		1	
KUHN	STRUCTURE SCIENTIFIC	1970		162		1	
KUHN	STRUCTURE SCIENTIFIC	1970		171		1	
KUHN	STRUCTURE SCIENTIFIC	1970		182		1	

Fig. 9 Citation counts of a book in web of science

Early in 2010, NLM started to introduce metadata for books and books chapters in PubMed. Besides NLM collects reference citations for the books in the digital collection called "Book Shelf". Here again, the E-Utilities of PubMed and the PMID can retrieve XML format for all the identifiers of the books or book chapters citing a given journal article.

The introduction of the h-index in bibliometric analysis implies obtaining the citation counts for each reference introduced in SERVAL. If the different providers are considered, there is only one single method to ensure the right match between the repository metadata and the external citation counts: the Web Service has to be queried by unique identifier (DOI, UT, PMID, PMCID). Some databases do accept queries combining other metadata elements (title + author name + year, journal name + volume + issue + start page, etc.). However this combination must be used only in absence of any identifier or in case the first method fails, because the chances of retrieving the right citation are fewer and the matching errors are not excluded.

In the bibliographic academic field, the most important identifier is certainly the DOI. Multidisciplinary, this identifier ensures the link to the electronic full text and it is largely included in many bibliographic databases. It can also be used as search criteria in most cases. The PMID, though limited to the biomedical field, is also widespread, and can be used as search criteria associated with the web service of WoS or SCOPUS. However, some systems like Citebase do not recognize it. As a matter of fact, the Open Archives or their aggregators accept the OAI-PMH protocol query "GetRecord". This method requires the internal identifier of the reference in the original archive so the papers archived in PubMed Central can be retrieved with the PMCID and not with the PMID. The PMCID is not included in the metadata of many databases or repositories and SERVAL is no exception. Yet, it is easy to obtain the PMCID for a given PMID on the fly, using one of the PubMed Web Services. A list of PMIDs can also be converted manually using the NLM "PMID to PMCID converter" web form (each query is limited to 2000 PMIDs).

The Web of Science Unique identifier (UT) is essential to extract data from this database with a guarantee of 100% match. The tests run in Lausanne revealed that sometimes, a reference included in Web of Science has no citation counts because the query by DOI or PMID gives no results. The same query by UT would have worked fine but at the moment, in the repository management system, only one external database identifier, in addition to the DOI, can be added. In most cases PMID was preferred.

Results

After some tests of the different databases and Web Services we reached the conclusion that only two resources could be used for the inclusion of the h-index in the bibliometric analysis performed in Lausanne: Web of Science and SCOPUS. However, the PubMed Central citation counts were also included in order to test this new data coming from Open Access documents.

Each SERVAL record and institutional author publications page has to be enriched with the citedness count coming from WoS and SCOPUS. Given that the majority of the biomedical records stored in the repository has the PMID and/or DOI, but is often deprived of the WoS identifier, the repository metadata with the citation counts have to be merged using the following simple search protocol:

1. If the SERVAL record has a UT, then it will be used to query Web of Science

2. For SCOPUS (and for Web of Science if the SERVAL record doesn't have an UT) use the query "[DOI] OR [PMID]"
3. If the SERVAL record doesn't have any external identifiers, then combine the fields "[Journal] AND [Volume] AND [Issue] AND [Start page]"
4. For PubMed Central, only the PMID is used in the query

Both databases offer a "light" web service returning the citation counts with the URL to build inbounds links

With the inspiration of some related projects like Bibliosight(30) and Socrates(31), and using the information shared by people who try to collect the citation counts to enrich the repository design (e.g. <http://hub.hku.hk/handle/123456789/44386>) or the link resolver main page of a document(32), we had designed a prototype using PHP Hypertext Preprocessor, a widely-used and general-purpose scripting language, combined with a web form allowing the choice between the 3 citation databases retained.

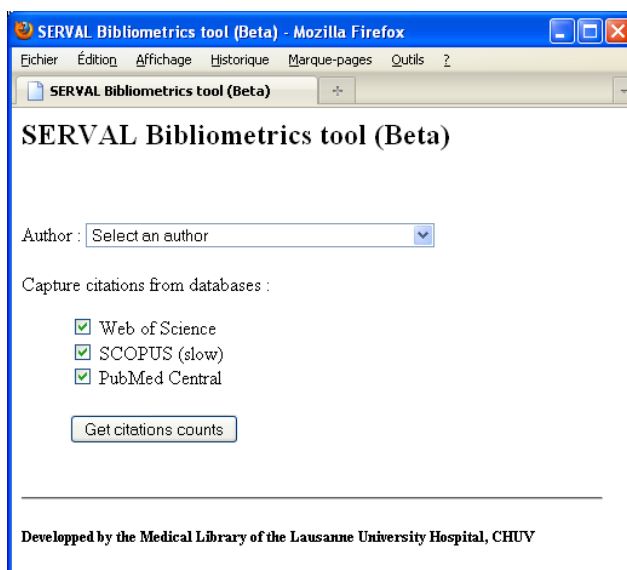


Fig 10 The Lausanne bibliometric prototype web form

The system, available at <http://www.bium.ch/bibliometrics/>, performs the following steps:

1. **Collect the publication list records metadata** for an author or a research unit from the institutional depository SERVAL
2. Use the **identifiers** or the **other metadata elements** to query each Web Service following the search protocol
3. **Parse** the responses and **extract the citations counts and URLs**
4. Display the **publication list with both citation counts** for each reference **and deep links**

The screenshot shows a web browser window displaying the UNIL website. The page is titled "Unisciences - UNIL - Jovan Mirkovitch". The main content area shows a list of publications by Jovan Mirkovitch, with citation counts from WoS, SCOPUS, and PubMed Central. The publications listed are:

- Richman L., Meylan P.R., Munoz M., Pinaud S., Mirkovitch J.**, An adenovirus-based fluorescent reporter vector to identify and isolate HIV-infected cells. *Journal of virological methods* 99(1-2), pp. 9-21, 2002. [PubMed]
SCOPUS@ Times Cited: 3
Web of Science@ Times Cited: 4
PubMed Central Times Cited: 1
- Fivaz J., Bassi M. C., Pinaud S., Mirkovitch J.**, RNA polymerase II promoter-proximal pausing upregulates c-fos gene expression. *Gene* 255(2), pp. 185-94, 09-2000. [PubMed]
SCOPUS@ Times Cited: 7
Web of Science@ Times Cited: 6
PubMed Central Times Cited: 3
- Fivaz J., Bassi M. C., Price M., Pinaud S., Mirkovitch J.**, Precisely positioned nucleosomes are not essential for c-fos gene regulation in vivo. *Gene* 255(2), pp. 169-84, 09-2000. [PubMed]
SCOPUS@ Times Cited: 2
Web of Science@ Times Cited: 2
PubMed Central Times Cited: 0
- Fivaz J., Chiara Bassi M. C., Pinaud S., Richman L., Price M., Mirkovitch J.**, EBV-derived episomes to probe chromatin structure and gene expression in human cells. *Methods in Molecular Biology* 133, pp. 167-82, 2000. [PubMed]
SCOPUS@ Times Cited: 1
PubMed Central Times Cited: 0
- Fivaz J, Bassi MC, Pinaud S, Mirkovitch J**, RNA polymerase II promoter-proximal pausing upregulates c-fos gene expression. *Gene* pp. 255, 185-194, 2000.
SCOPUS@ Times Cited: 7
- Fivaz J, Bassi MC, Price M, Pinaud S, Mirkovitch J**, Precisely positioned nucleosomes are not essential for c-fos regulation in vivo. *Gene* pp. 255, 169-184, 2000.
- Souici A.C., Mirkovitch J., Hausel P., Keefer L.K., Felley-Bosco E.**, Transition mutation in codon 248 of the p53 tumor suppressor gene induced by reactive oxygen species and a nitric oxide-releasing compound. *Carcinogenesis* 21(2), pp. 281-287, 2000. [PubMed]
SCOPUS@ Times Cited: 27
Web of Science@ Times Cited: 30
PubMed Central Times Cited: 4
- Price M., Fivaz J., Jotterand A., Mirkovitch J.**, Tissue-specific chromatin structure at the

Fig. 11 The SERVAL publication list enhanced with WoS, SCOPUS and PubMed Central citation counts and deep links

5. **Create a new dataset taking the highest citation counts for each publication**
6. **Calculate several bibliometric indicators for each database and the new dataset:**
 - Number and percentage of references retrieved in the database
 - Total sum of the times cited
 - Average of citations per retrieved article
 - Number of publication never cited (citation count = 0)
 - **h-index**
 - **g-index and g/h ratio**
7. **Generate a table including the bibliometric indicators and the complete list of publication identifiers retrieved from different databases and sorted by the highest citation counts.**

UniSciences Université de Lausanne | Annuaire | Actualités | Aide | Login

Langues: fr | en MyUNIL

RECHERCHE

AFFICHER LES DONNÉES à

Données actuelles

Unisciences en bref Jovan Mirkovitch

Liste des personnes

Liste des unités

Aide

Coordonnées Enseignements Publications **Bibliométrie**

Number of publications in SERVAL: 31

Metrics	Web of Science®	SCOPUS®	PubMed Central	Max.
Publications retrieved	24 of 31 (77.42%)	28 of 31 (90.32%)	25 of 31 (80.65%)	29 of 31 (93.55%)
Sum of the Times Cited	2103	912	510	2124
Average Citations per Item	87.63	32.57	20.40	73.24
Number of never cited publications	0 of 24 (0.00%)	0 of 28 (0.00%)	5 of 25 (20.00%)	0 of 29 (0.00%)
h-index	16	13	10	16
g-index	45	30	22	46
g/h ratio	2.81	2.31	2.20	2.88

Details :

Ref.	SERVAL ID	Year	DOI	PMID	UT	Cited by				Average citations per year			
						WoS	Scopus	PMC	Max.	WoS	Scopus	PMC	Max.
1	BIB_CC4FC6004183	1984		6091913	A1984TS61800024	897	391	225	897	33.22	14.48	8.33	33.22
2	BIB_EA0C2B8516BC	1991		1901265	A1991FE95000023	296		70	296	14.80		3.50	14.80
3	BIB_F4DB202D41B1	1988		3132558	A1988M424800010	125	58	33	125	5.43	2.52	1.43	5.43
4	BIB_3C16F94535C1	1988		3132557	A1988M424800009	122	54	25	122	5.30	2.35	1.09	5.30
5	BIB_BB24B8484DC5	1989		2726767	A1989U940700041	96	24	25	96	4.36	1.09	1.14	4.36
6	BIB_F74E14052B12	1994		7523862	A1994PM68400026	81	81	24	81	4.76	4.76	1.41	4.76
7	BIB_55255A09885C	1993		8375375	A1993LW53600001	80	61	10	80	4.44	3.39	0.56	4.44
8	BIB_8AD0CC74260F	1986		3098982	A1986D342500010	63	20	15	63	2.52	0.80	0.60	2.52
9	BIB_3FC72A7C088A	1992		1729591	A1992GW05300001	59	35	22	59	3.11	1.84	1.16	3.11
10	BIB_27C6980EFC77	1987		2894689	A1987L347900012	57	23	10	57	2.38	0.96	0.42	2.38
11	BIB_BA44BB657361	1991		1989908	A1991EV40900009	56	16	15	56	2.80	0.80	0.75	2.80
12	BIB_8175F72EAF42	1998	10.1006/imbi.1998.1905	9671550	000075177600003	32	32	13	32	2.46	2.46	1.00	2.46
13	BIB_14120	2000		10657969	000085503400023	30	27	4	30	2.73	2.45	0.36	2.73
14	BIB_05D2ECF56948	1995		7554056	A1995RV29600010	28	26	2	28	1.75	1.63	0.13	1.75
15	BIB_CBD09F7B3C09	1992		1384813	A1992JU06900004	19	10	8	19	1.00	0.53	0.42	1.00
16	BIB_E797B5424FC1	1988		2968256	A1988N798700008	18	5	2	18	0.78	0.22	0.09	0.78
17	BIB_891FA46D520A	1997		9442399	000070994500028	10	8	0	10	0.71	0.57	0.00	0.71
18	BIB_33AF2FE4452B	1989		2570675	A1989AR04200029	9	2	0	9	0.41	0.09	0.00	0.41
19	BIB_4E298AE65132	1992		1325913	A1992JM09700039	9	3	2	9	0.47	0.16	0.11	0.47
20	BIB_834	1997					8		8		0.57		0.57
21	BIB_17149	2000					7		7		0.64		0.64
22	BIB_378D3ED4BCC7	2000		11024278	000089777600005	6	7	3	7	0.55	0.64	0.27	0.64
23	BIB_97E12DDFDA33	2002		11684299	000173059200002	4	3	1	4	0.44	0.33	0.11	0.44
24	BIB_B79909288B7D	1995		8531716	A1995BE08E00039	2	3	0	3	0.13	0.19	0.00	0.19
25	BIB_9CB0963CB63B	1998		9573349	000073572000018	2	2	1	2	0.15	0.15	0.08	0.15
26	BIB_856063B40714	2000		11024277	000089777600004	2	2	0	2	0.18	0.18	0.00	0.18
27	BIB_9236	1998					2		2		0.15		0.15
28	BIB_1252B74C8D23	2000		10561838			1	0	1		0.09	0.00	0.09
29	BIB_14F8513CCA30	1993					1		1		0.06		0.06
30	BIB_9105	1998											
31	BIB_17147	2000											

Contact - Copyright

Unicentre - CH-1015 Lausanne - Suisse - Tél. +41 21 692 11 11 - Fax +41 21 692 20 15

Fig. 12 The SERVAL publication list converted in a bibliometric table "on the fly"

The advantage of mixing data from WoS, SCOPUS and PubMed Central is that we can take the highest citedness score per record and obtain a new kind of metric less sensible to specific database errors and shortcomings (10, 33).

Conclusion and Future work

The Web Services offered by reference-enhanced databases are particularly interesting when combined with the accuracy of metadata and the comprehensiveness found in the institutional repositories. If the Impact Factor or other journal centered metrics usually change on a yearly basis only, the bibliometric information like citation counts changes each week and must be renewed. This situation fosters the mash-up techniques and the merging of metadata on demand. It allows introducing the h-index in the bibliometric research assessment process.

Higher standardization in the access to the data providers (at the moment only the Open Archives uses a standard protocol, the main databases like Web of Science, SCOPUS and PubMed have their own system and API) and the increase of the bibliometric information available and delivered in a machine readable format (XML or JSON) would improve the efficiency and the interest of the Web Services use.

Currently, rich and accurate metadata can only be retrieved using unique identifiers like DOIs, PMIDs and UTs. Those identifiers are the essential pivot between bibliographic databases, Open Archives and third part Web Services. It is very important to collect them as soon as possible and as many as possible in the institutional repositories and library catalogs.

Comparing all the bibliometric resources and implementing a technical solution extracting the citations counts and making a mash-up with the repository metadata was not easy, if we consider that this is a field experiencing big changes with a quickly moving technology. Within one year from now, the landscape of the reference-enhanced databases will certainly be different and we must be awake and refresh our bibliometric system using the better resources and tools to improve the research assessment process in our faculty.

The next step is to improve the prototype and to adapt it to work with other sources of metadata like PubMed with the purpose to make bibliometric analysis of external authors. The inclusion of an internal Web Service returning the journal IF for a given ISSN will also be tested.

The field of researchers' identifiers develops also very quickly. New possibilities have to be explored, particularly those offered by the platforms like ResearcherID (<http://www.researcherid.com>) or ORCID (Open Researcher and Contributor ID, <http://www.orcid.org>) and the emerging Web services specialized in names like the VIAF project (<http://www.viaf.org>) or the Wikipedia API (<http://www.mediawiki.org/wiki/API>). The techniques of disambiguation and retrieval of synonyms could be really important outside the repository sphere, where the lists are supposed to be "homonymous free".

At last, we suggest the inclusion of a new metric taking for each reference the highest citedness score extracted either from WoS, Scopus or PubMed Central. This would be a **mixed data h-index**. Anyhow the system is flexible and open to the inclusion of new metrics and techniques of normalization discussed with the research evaluation unit.

Bibliography

1. The Thomson Reuters Impact Factor [This essay was originally published in the Current Contents print editions June 20, 1994, when Thomson Reuters was known as The Institute for Scientific Information® (ISI®)]. Available from:
http://thomsonreuters.com/products_services/science/free/essays/impact_factor/.
2. Schwartz S, Hellin JL. Measuring the impact of scientific publications. The case of the biomedical sciences. *Scientometrics*. 1996;35(1):119-32.
3. Hirsch JE. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*. 2005;102(46):16569-72.
4. de Kaenel I, Iriarte P. Efficacy and benefits of web services for metadata acquisition: an overview based on Swiss institutional repositories [poster]. CERN Workshop on Innovations in Scholarly Communication (OAI6); Geneva, Switzerland: CERN; 2009.
5. Lausanne University. Déclaration d'affiliation : Directives UNIL pour déclarer l'affiliation des chercheurs dans les publications et communications scientifiques. Available from:
http://www.unil.ch/fbm/page39911_fr.html.
6. SCImago Journal & Country Rank: Journal Rankings. Available from:
<http://www.scimagojr.com/journalrank.php>.
7. Leydesdorff L, Opthof T. Scopus's Source Normalized Impact per Paper (SNIP) versus a Journal Impact Factor based on Fractional Counting of Citations. *Journal of the American Society for Information Science & Technology*. Forthcoming 2010.
8. Centre for Science and Technology Studies (CWTS). CWTS Journal Indicators. Available from: <http://www.journalindicators.com/>.
9. Garfield E. The evolution of the Science Citation Index. *Int Microbiol*. 2007;10(1):65-9.
10. De Bellis N. *Bibliometrics and citation analysis : from the Science Citation Index to cybermetrics*. Lanham, Md.: Scarecrow Press; 2009.
11. Egghe L. Theory and practise of the g-index. *Scientometrics*. 2006;69(1):131-52.
12. Thomson Reuters. Web of Knowledge Quality and Quantity: Real facts, Real numbers, Real Knowledge. Available from: <http://wokinfo.com/realfacts/qualityandquantity/>.
13. Thomson Reuters. New! Web of Science® Web Services. Available from:
http://wokinfo.com/products_tools/products/related/webservices/
14. Thomson Reuters. Web of Science Article Match Retrieval. Available from:
<http://researchanalytics.thomsonreuters.com/solutions/amr/>.
15. Jacsó P. The pros and cons of computing the h-index using Web of Science. *Online Inf Rev*. 2008;32(5):673-88.
16. Jacsó P. As we may search - Comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases. *Curr Sci*. 2005;89(9):1537-47.

17. Elsevier. Scopus Content Coverage Guide. Available from: <http://info.scopus.com/scopus-in-detail/facts/>.
18. Elsevier. Scopus API. Available from: <http://info.scopus.com/scopus-services/integration/solutions/api/>.
19. Eaton A. Scopus API. In: HUBLOG. 2007 July. Available from: <http://hublog.hubmed.org/archives/001512.html>.
20. Jacsó P. The pros and cons of computing the h-index using Scopus. *Online Inf Rev.* 2008;32(4):524-35.
21. NIH's National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM). Entrez Programming Utilities. Available from: http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html.
22. NIH's National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM). PMC Utilities. Available from: http://www.ncbi.nlm.nih.gov/pmc/about/PMC_Utilities.html.
23. Citebase Search. University of Southampton. Available from: <http://www.citebase.org>.
24. Ridge B. Google Scholar Firefox Add-on. Available from: <https://addons.mozilla.org/en-US/firefox/addon/10310/>.
25. Harzing AW. Publish or Perish. Available from: <http://www.harzing.com/pop.htm>.
26. Jacsó P. The pros and cons of computing the h-index using Google Scholar. *Online Inf Rev.* 2008;32(3):437-52.
27. CrossRef History/Mission. Available from: <http://www.crossref.org/01company/02history.html>.
28. CrossRef Forward Linking. Available from: http://www.crossref.org/02publishers/forward_linking_howto.html.
29. Jacsó P. Citation-enhanced indexing/abstracting databases. *Online Inf Rev.* 2004;28(3):235-8.
30. Bibliosight. JISC; Available from: <http://www.jisc.ac.uk/whatwedo/programmes/inf11/jisc/bibliosight.aspx>.
31. Owens R, Mast NG, Glance DG, McEachern D. Making the Most of Citation Data: The Integration of Thomson Reuters Web of Science and UWA's Research Management System, Socrates Biometrics Conference; Brisbane, Australia: ISI Web of Knowledge Australia; 2009.
32. vLib project team in the Max Planck Digital Library (MPDL). New in MPG/SFX: "View this record in Web of Science". In: Max Planck vLib News. Available from: <http://blog.vlib.mpg.de/new-in-mpgsfx-view-this-record-in-web-of-science/>
33. Jacsó P. Database source coverage: hypes, vital signs and reality checks. *Online Inf Rev.* 2009;33(5):997-1007.