High Energy Physics Libraries Webzine

Issue 4 / June 2001

Torii, an Open Portal over Open Archives

Sara Bertocco

25/05/2001

Abstract

The world of academic publishing is undergoing many changes. Everywhere paper-based publishing is being replaced by electronic archives and ink printing by bits. Unrestricted (web) access to many resources is becoming a fundamental feature of the academic research environment. Particularly in the high-energy physics community, the pre-print distribution has moved completely away from the paper-based system into a fully electronic system based on open archives. At the same time, freely accessible peer-reviewed journals have started to challenge the more traditional, and paper-based journals showing that the entire paper-based cycle can be effectively replaced by a web-based one. The TIPS project was born in this environment and from these observations. It is based on the idea that further progress in information distribution and scientific publishing on the web requires some key ingredients: the implementation of a more extensive semantic structure in the documents that are exchanged; a unified, desktop-like, web access to different archives, journals and services to manage information; the availability of better information retrieval and filtering techniques. TIPS is part of the European Union Fifth Framework Program Information Society Technologies Program: IST-1999-10419, and its first implementation is a portal named *torii.* At present, torii is undergoing an evaluation phase, so it is available only for a restricted test user group.

Introduction

Bits vs. Paper. The world of academic publishing is undergoing many changes. Everywhere paper-based publishing is being replaced by electronic archives and ink printing by bits. Unrestricted (web) access to many resources is becoming a fundamental feature of the academic research environment, as shown by the pre-print and genome databases. Traditional publishing houses have been slow in understanding this shift and are only now reacting accordingly. They are unwilling to switch to a new model, especially because it is still unclear how profit can be made from the web publishing market. So far, the compromise has been to charge the same subscription rates as in the past while throwing in web access to some of the same information that is sold on paper. This is not going to work for two reasons:

- access to basic information is recognized as a common good not to be negotiated or paid for. Accordingly, more and more of the same information is provided freely on the web and charging for it has become very unpopular;
- very often the web version offered contains almost nothing more than what is available for free or on the paper version and is therefore rarely accessed, if at all.

Copyright on Basic Information. The form of copyright transfers from authors to publishers in academic publishing is changing accordingly: almost every publisher tolerates the distribution of material in separate archives out of their control, either local to institutions, or worldwide. Hence the copyright on basic information is actually returning to the hands of the scientific community, through open archives. Copyright coming from the publisher-added values---which are selection, printing and distribution---is no longer justified since:

- selection is done by the scientific community,
- centralized paper printing is unnecessary, and
- distribution is virtually cost-less.

Moreover, it is useful to bear in mind that it has become increasingly difficult to prevent file sharing among groups of users, thus making it unpopular but also impractical to charge for direct document access (see the example of watermarking in the world of commercial music distribution).

An Opportunity and a Challenge. The vacuum left behind by traditional publishing houses offers a great opportunity to the academic world to resume control over the process of publishing its own results. After all, this is the way it used to be up to the second half of this century when commercial enterprises took over the job from the scholarly societies. Leading the way is the high-energy physics community where the pre-print distribution has moved completely away from the paper-based system into a fully electronic system based on open archives.[1] At the same time, freely accessible peer-reviewed journals have started to challenge the more traditional, and paper-based journals showing that the entire paper-based cycle can be effectively replaced by a web-based one.[2]

New Tools: Hardware vs. Software. It seems clear that some of the most decisive future changes are going to take place in hardware development. The best example is paper. For years we have been struggling with computer screens, the problem being that it is difficult to really concentrate when reading a screen. On the other hand, paper, as beautifully apt to carry written text and pictures as it is, lacks the advantage of computer screens. It now seems likely that in the near future there will be a kind of paper in which ink dots, or e-ink, will be controlled by chips thereby providing the advantages of both media. Another example is high-quality printing (and binding) machines. The day all libraries will be equipped with such machines the exchange of papers will take place entirely on-line and mailing of journals will stop altogether. A final example is 3rd-generation technology in which mobile phones, TV and web browsing come together, raising the issue of providing access to, for instance, hand-held devices with minimal operative systems. It is therefore important to bear in mind that many parts of the software we are going to write today will be implemented tomorrow on new and better hardware, but also on a greater variety of hardware types (e.g., existing mobile phones browsing the web).

What to Change and What to Preserve. Coming to information production and dissemination within academia in general and the high-energy physics community in particular, there are things that are done the way they are done out of habit (shaped by the means available) and there are things that are done the way they are because it is the best way. As for the former, we can try and change them, whereas the best must simply be preserved and translated into the new media. Of course, the real art in innovating consists in understanding which is which, and for this reason user

requirements have been carefully analysed. While there are many things that clearly belong to the former class like the means for writing, sending around and storing (library stacks) scholarly papers, other features are more difficult to judge. As an example, consider the present dual structure where preprints co-exist side by side with journals. Is this an artifact of the slowness of traditional journals or a useful distinction? Similarly, should the present procedure of review remain or do we have different options? Another important area of discussion is whether to centralize all documents in few repositories or let each author post them on his/her own web site and have the services provide distributed linking.

The TIPS Project

The TIPS Project: the Basic Idea. The TIPS project is based on the idea that further progress in information distribution and scientific publishing on the web requires

- the implementation of a more extensive semantic structure in the documents that are exchanged,
- a unified, desktop-like web access to services and tools, to manage information, and
- the availability of better information retrieval and filtering techniques.

The project addresses these problems in the specific environment of academia, in particular the high-energy physics community, where they were first put to work and where we hope to find a receptive test-bed for the architecture and tools we want to introduce. It is no historical accident that the marriage of hypertext and Internet known as the web was first designed to support science publishing and information exchange at *info.cern.ch*, the high-energy physics laboratory near Geneva. Advanced research in science has often provided new technical tools that are then taken up by industry and made available to every one of us, and the web has been just one of them. The TIPS project has been financed by European Union's Fifth RTD Framework Program (1998-2002) within the Information Society Technologies Program (IST) and is carried out by a consortium of six partners: SISSA and the University of Udine (Italy), City University (UK), UJF (France), CERN (Switzerland) and IoPP (UK).

Multilayered Documents. The overall architecture of the TIPS project is based on the concept of the multilayered document and its dynamic access. A multi-layered document is a document along with the full set of annotations, comments, changes and additions that have been made to it.[3] As an example, think of a scholarly paper. The original text is written by the author and submitted to a public archive to spread its contents and to a journal for review. After this stage, the paper will be revised, comments will be added on top of it and, if accepted, a judgment on its acceptability will be attached to it as it is published in a recognized journal (this will contain both the referee's opinion and the journal clearance stamp). Any seminars that the author gives in the meantime at conferences are additional (multimedia) layers traceable back to the original document. Hence, the multi-layered document is a stack of documents that we want to manipulate. Technically, it could be an entry in a database, and as new layers are added so the entry column is modified in the database, or it could be a collection of documents managed by a web server that keeps track of their relationships and modifications. Consider the second case, you will have access to the documents and be able to modify them on the fly (for instance, by attaching a little yellow stick-on-like note with a comment). Other people will also have access to the original document and to your comment and see the structure (and verify who has done what through electronic signatures). This is the other side of the web as it was originally conceived [4], but which has not yet come into being: the real web of knowledge mimicking our real life interactions.

Peer Review and Quality Control Tools. Applying the multilayered document model, the TIPS project aims to provide a set of quality control tools that will help in setting up a truly peer reviewed

system in which the community as a whole will participate in accruing comments on a given document. It is easy to set up a system in which for any document accessed the user is also offered the possibility of recording comments and judge the document itself, thus creating a continuously updated database of annotations on top of the original document. Access to this database will automatically give a ranking of documents much richer than the current published/unpublished system. A more sophisticated system can also be implemented. See, TIPS documentation at [5]

Dynamic Access. Access to a multilayered document must be dynamic. According to who you are at a given moment---reader, author, referee, editor---you have access to different layers. As an example, consider a web page that recognizes you as you access it and presents all the possible information and tools you asked to find there. Dynamic access requires an appropriate interface between the multilayered documents and the users. It also requires intelligent agents to sift through the increasingly large amount of information to shape it into some hierarchy, thus making it usable by the user.

Filtering Tools. As the amount of data grows every day, it is necessary to filter part of the information reaching the users according to profiles either set up by the users themselves (cognitive filtering) or by proximity of interests among user groups (social filtering). This can be implemented in a service providing access to a daily-updated archive in which the listing of the records is filtered by the system, for instance by ranking the documents by the potential relevance for the user according to his/her profile, or to the profile of the user group he/she belongs to. According to the same profiles, the system can also suggest potentially interesting documents that are in other archives and that would have otherwise escaped the user's notice.

Assisted Search. Search engines are necessarily the core of any system providing access to stored documents. The workings of most current search engines on the web is essentially known as the *grep* command on Unix platforms and it therefore cannot distinguish, for instance, between a book written by Anthony Trollope, a book about the writer Anthony Trollope and a book about a book by Anthony Trollope (assuming that you have put the first name into the search query and are not being presented also with information on Joanna Trollope). Adding semantic structure to the web means that such a distinction is made and can be used by the software. The current web is still far from being able to accommodate a similar scenario, because it lacks the necessary semantic structure. However, things are changing fast and the semantic web is becoming a reality. TIPS's main search is powered by Okapi, a probabilistic full-text engine. In order to improve the success rate of searches, an assistant has been added. The assistant tries to prevent dead-end searches by suggesting terms (from a terminological thesaurus), and extricates users from the one-thousand-items search results.

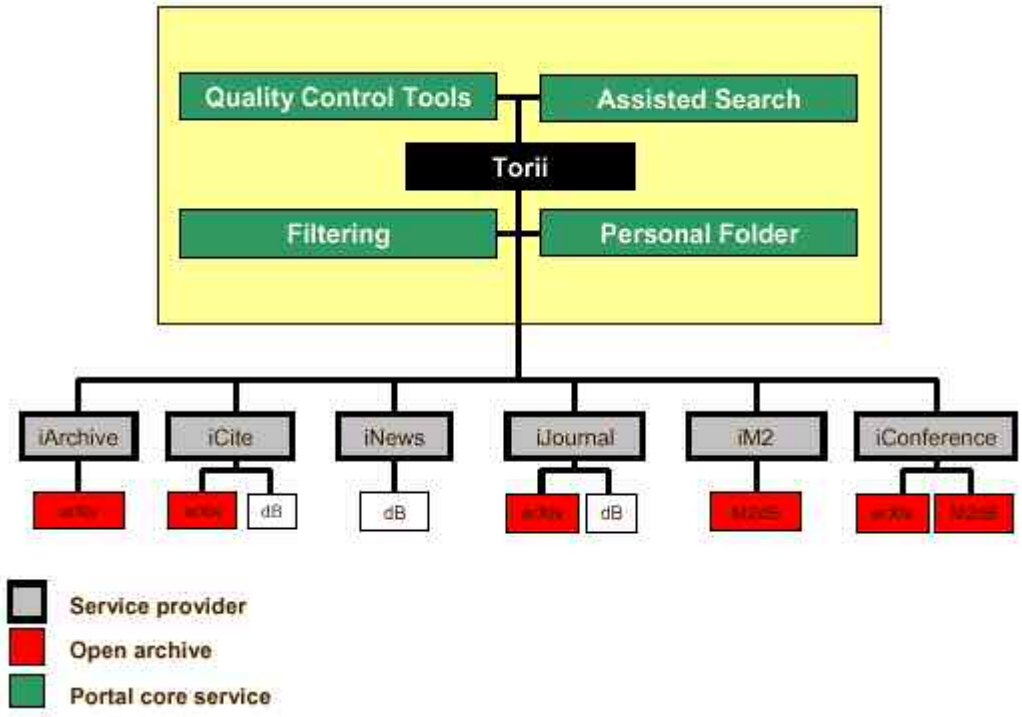
The Torii Portal

A Possible Scenario: the Full Portal. The first prototype of the project TIPS is a portal named torii (after the Shinto gate that marks the boundary between the temple and the everyday world). This portal provides end users with a single gateway to personalized information and comprehensive support for their work, giving access to the tools they need in their everyday work. Ideally such a system replaces the user's desktop environment. Main features of torii are the following:

- it integrates dynamic access in a wide variety of data formats, allowing users to share, manage, maintain information from one central user interface (comprehensive);
- it organizes access to data, without storing the data itself (location-transparent);
- it organizes access to information for users to browse, i.e. using impact factors to order the browsing of a list of documents (organized);

- it uses acquired data and information for further processing and analysis, using for example user profiles to organize the browsing of documents or to suggest to the user other equally interesting articles (filtered);
- it assembles personalized views of key information and notifies users on the availability of new material, using a personal folder in which to place suggested material and from which to evaluate user preferences (personalized);
- it supports extensions for new types of information, implementing a unified interface for open archives (extensible);
- it automatically identifies and organizes access to new content, providing selective listing and browsing mechanisms (automated).

Torii Architecture. Torii implements the three-tiered architecture shown on Fig.1. On the bottom, there is a first tier which consists of resource providers, i.e. a set of open archives and databases. In the middle there is a second tier of services, that are all external to the portal, having their own web interface, and each using its own database and/or open archive. These services are integrated within the portal if they implement a set of specifications that makes the integration possible. On the top is torii which integrates the underlying services with a set of tools to manage information harvested from open archives and databases.



XML, Open Archives and Databases. Key features for the integration of dynamic access to the information into the portal are the XML language and the open archive initiative protocol. The XML language is used to encapsulate the exchanged information, originally stored in a variety of formats, in a common XML structure. This XML metadata structure will represent the semantic aspects related to the data. For example, if an article is stored with a record

```
<article>
    <title>      ...   </title>
    <author>     ...   </author>
    <abstract>   ...   </abstract>
            ............
</article>
```

it will be possible to perform the kind of semantic search as described above. The Open Archives Initiative Protocol is the basic communications protocol between the portal and

the underlying services, operating in a location-transparent way. The Open Archives Initiative, indeed, develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content. The goal of the open archives initiative protocol for metadata harvesting is to supply and promote an application-independent interoperability framework that can be used by a variety of communities engaged in publishing content on the Web [6]. This is precisely the aim of torii. Through the use of the Open Archive Initiative Protocol, torii will be easily extensible to each archive that will implement the Open Archives Protocol. At the moment, at the base of the portal there are two open archives: mmdb [7], which is an archive for multimedia documents built at SISSA, and the Los Alamos arXiv.

Services. On the middle tier of torii, in the current version, there are three services:

- iArchive: a service working on arXiv and powered by-
- iCite: a citation harvesting system built at SISSA and based on ResearchIndex, a citation indexing service of the NEC Research Institute. iCite is able to provide the impact factors of the documents stored in arXiv;
- m2db: a service on the multimedia archive that allows you to upload, download, and retrieve multimedia documents through an interface implementing the Open Archive Initiative Protocol.

The original TIPS project also included other services like a conference handling system (for the complete organization of a conference), a job search/offer server, etc. If someone will implement these services with the related archives according to the open archive initiative specifications, it will also be possible to integrate them into the portal.

Torii. Torii will give access to the open archives and all subscribed services by means of a single, integrated desktop environment. The torii interface will be the place where all information, tools and services required by the user in his/her daily work will be available. Torii will give access to the open archives and all subscribed services by means of a single, integrated desktop environment. For services that require identification, a user identifier and password will be automatically negotiated, with information filtered and organized by the system. Moreover, the integrated system will be able to exchange information among users and therefore improve collaboration and information sharing. Actually, torii includes

- a set of quality control tools that allow users to express judgements about documents collected using the system, filling predefined forms;
- a set of cognitive and social filtering tools based on analysis and elaboration of personal profiles defined by the users (the figures that follow, for example, show the form used to define a personal profile -Fig.2- and the mechanism for adding information to the personal profile by indicating an interesting document -Fig.3-);
- a mechanism for organizing the browsing of information, not only by date or by document identifier (the most common ways), but also using results of filtering processes or impact factors (evaluated by the iCite service);
- a personal folder where users can store and organize the information pertinent to their work.

The torii environment also comprises a set of assisted search tools. The integration work is in progress.


Netscape: TIPS Project
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://torii.sissa.it:11000/torii/auth/index.jsp What's Related
Internet Lookup New&Cool
Torii
Hello, fabio. Welcome back to Portal 0.5.
Monday, 4 June 2001
Get Browse Search Log-out Personalize
Modify Account
Profile
Personal folder
Managment User Profile(s)
In Box
Interesting Documents
Archived Documents
Trash!
Profile
Name:
Description:
Keywords
Interesting text
Interesting Document
(Input the complete identifier, e.g. arXiv:hep-th/0103118)


Netscape: TIPS Project
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://torii.sissa.it:11000/torii/auth/index.jsp What's Related
Internet Lookup New&Cool
Torii
Hello, fabio. Welcome back to Portal 0.5.
Monday, 4 June 2001
Get Browse Search Log-out Personalize
Archive
M2db
arXiv sector: hep-th
Browse
2001 03
List
last 24 hours
List
Personal folder
In Box
Interesting Documents
Archived Documents
Trash!
arXiv:hep-th/0006184
A Brane World Model with Intersecting Branes
Author: Pavsic, Matej
Abstract: A brane world model is investigated, in which there are many branes that may intersect and self intersect. One of the branes, being a 3-brane, represents our spacetime, while the other branes, if they intersect our brane world, manifest themselves as matter in our 3-brane. It is shown that such a matter encompasses dust of point particles and higher dimensional p-branes, and all those objects follow "geodesics" in the world volume swept by our 3-brane. We also point point out that such a model can be formulated in a background independent way, and that the kinetic term for gravity arises from quantum fluctuation of the brane.
Submitted: 2000-06-23
Paper: Source, PostScript, PDF References, Citations
Add to Personal folder Evaluate this document Feedback on profile
Netscape: Question
By saying OK, you add
arXiv:hep-th/0006184
to your profile
OK Cancel

Economic Model

Open Archives and Services. The lesson from the open-source experience is that it is possible to maintain free access to valuable information (program sources in that case), while charging for services built on top of this information. The same approach can be translated into academic publishing by separating the publishing into three layers:

- archives, where the files of papers (or proceedings, or multimedia documents, possibly even books) are kept and to which free web access is granted;
- services, built on top of these archives, the access to which may be charged for; and
- portals, where the services are integrated and new tools implemented.

The economic model of the portal, similar to that of single services, can be based on a voluntary-based membership by libraries in which access is unrestricted but subscription is encouraged (on the model of many public museums in the US) and gives the right to additional services. The extent of these additional services will depend on individual cases. This structure gives, at the same time, unrestricted access to the basic information and justifies charging for costs incurred for setting up additional (and more refined) services above the open archive layers. Archives are going to be run by institutions ready to pay these (limited) costs in exchange the prestige and influence that come from it. Owning an archive is going to be the same as having a Nobel-prize winner in the Faculty. The cost of running such an archive will be paid back by the increased prestige and, accordingly, by the larger number of students and a greater availability of grants. On the other hand, it will be possible to charge for the services because they represent added value. They will be run by anyone willing to take up the job.

Scholarly Journals. The example of scholarly journals is helpful in understanding the details of the doubled tiered structure between open archives and services. It is now practically within reach for any middle-sized institution to start an electronic journal, that is a journal published only on the web. This entails a dramatic reduction in costs and is mainly due to three changes that have taken place in the publishing procedure:

- authors do their own typesetting,
- the editorial work can be run by software robots, and
- there is no need to print the journal on paper.

On the other hand, some costs have remained the same, namely, editorial review fees (if required) and copy-editing (if desired). Barring the last two, only a modicum of secretarial work is left to be covered by the publishing institution.(Of course, there are hidden costs, like internet connection charges and system manager salaries to be taken into account in the overheads). The journal can be run in two different ways: either as a service on top of the open archive of pre-prints, or as a parallel open archive. In the former case it will charge users for accessing the added information (and it will belong to the second tier), in the latter case it will give free access to its archive (and it will belong to the first tier).

Libraries as Service Providers. In the (very close) future, the role of libraries, as paper archives are progressively replaced by digital ones, will be to

- maintain the open archives (probably from consortia of institutions and libraries), and
- provide access for its users to high-level services by subscribing to them.

Aside from collections of old paper resources, new documents accessed through archive and service providers will be available on screen and, on request, will be printed locally by the new generation digital printing/binding machines. The budget nowadays spent in journal subscriptions will progressively be transferred to service subscriptions, archive maintenance and hardware for users. It is not difficult to envisage a researcher browsing the listing of the most recent preprints on a hand-held device while driving to his/her workplace (the service having already selected those potentially of interest for him/her), selecting some of them while waiting for a green light, having them down-loaded via an infrared connection while walking by a high-quality printer in the school library, and finding them already bounded next to his/her office.

Conclusion

The portal torii is a prototype of the TIPS project that we will exploit to perform the first usability tests. The prototype is now available for a test user group, that will use the portal, test it and express judgments on its usability, utility, and validity in general. The results of the tests will suggest to the developers what is well done and what needs to be changed. On the basis of the user tests, there will be a revision of the system specifications and a new portal release that will be available to the whole user community in the middle of 2002.

Acknowledgements

I would like to thank all the people involved in the realization of the portal torii for their contribution to the present work: Fabio Asnicar, Loriano Bonora, Marco Fabbrichesi, Fabrizio Nesti, and Cristian Zoicas, *from SISSA, Trieste (Italy)*; Massimo Armellini, Giorgio Brajnik, Massimo Di Fant, Stefano Mizzaro, and Carlo Tasso, *from Università di Udine, Udine (Italy)*; Susan Jones, Murat Karamuftuoglu, Stephen E.Robertson, Fabio Venuti, and Xinkun Wang, *from City University, London (UK)*; Catherine Berrut, Marie-France Bruandet, Jean-Pierre Chevallet, and Nathalie Denos *from Université Joseph Fourier, Grenoble (France)*; Michel Goossens *from CERN, Geneva (Switzerland)*; Nigel Hollingworth *from IOP, Bristol (UK).*

References (sample only)

1. arXiv.org e-Print archive
   URL: <http://arXiv.org/>
2. the journal of high energy physics
   URL: <http://jhep.sissa.it/>
3. UC Berkeley Digital Library Project
   URL: < http://galaxy.cs.berkeley.edu/info/>
4. Weaving the Web
   URL: <http://www.w3.org/People/Berners-Lee/Weaving/>
5. TIPS documentation
   URL: <http://tips.sissa.it/>
6. Open Archive Initiative
   URL: <http://www.openarchives.org/>
7. MultiMedia DataBase
   URL: <http://mmdb.sissa.it/>

Author Details

Sara Bertocco
SISSA - ISAS
Scuola Internazionale Superiore Studi Avanzati di Trieste
via Beirut, 2-4
34014, Trieste
Italy

***Tel**: +39 040 3787513*
***Email**: bertocco@medialab.sissa.it*

Sara Bertocco is employed as Software Engineer at SISSA - ISAS (the International School for Advanced Studies of Trieste, Italy). She works on design and development of software systems for scientific communication.

For citation purposes:
Sara Bertocco, "Torii, an Open Portal over Open Archives", High Energy Physics Libraries Webzine, issue 4, June 2001
URL: <http://library.cern.ch/HEPLW/4/papers/4/>