International Congress of Mathematicians 2002 – Beijing

EIC-satellite conference at Tsinghua University
Electronic Information and Communication in Mathematics
Beijing, Aug. 28-31, 2002

Mathematics Subject Classification and related schemes in the OAI framework

Antonella De Robbio, Dario Maguolo
Mathematics Library – University Library System
University of Padova – ITALY

Alberto Marini
Institute for Applied Mathematics and Information Technology
National Research Council (CNR-IMATI), Milano - ITALY

Abstract

This paper aims to give a feeling of the roles that discipline-oriented subject classifications can play in the Open Archive movement for the free dissemination of information in research activities.
Mathematics, and Mathematics Subject Classification, will be the focuses around which we will move to discover a variety of presentation modes, protocols and tools for human and machine interoperability.
The Open Archives Initiative (OAI) is intended to be the effective framework for such a play.
In the first part of this paper, we start by describing the most important subject classification schemes in mathematics and related disciplines. Then we sketch the structure of discipline-oriented schemes in view of browsing and we give an account of different browsing modalities, implemented in the tools we produced and collected in The Scientific Classifications Page. Finally we give an insight on the design, implementation and use of a programming language for the generation of hypertextual presentations of complex structured data.
In the second part, we list different strategies for e-print communication in scientific research, up to the basic definitions of the Open Archives Initiative.
A review of the functionalities actually implemented in OAI compatible archives managed by the EPrints software will lead us to some working hypotheses about the roles that subject classifications in mathematics and related disciplines can play in the scenarios of the Open Archives movement.

Contents

1 – Subject classification schemes

Subject classification schemes are primary tools for the organization of knowledge and terminology in scientific disciplines.
They are produced mainly by professional societies, or academic and research institutions, often to be employed in their own bibliographic databases. Although many of the issuing bodies have national or regional scope, subject classification schemes are generally international in scope, and are intended to be a communication tool for the international scientific community.

1.1 – Schemes for mathematics

Mathematics Subject Classification (MSC) [1] is developed by the editorial offices of the two world's most important bibliographic databases for mathematical research:
- *MathSci*, which is produced by the American Mathematical Society, and
- *Zentralblatt MATH*, which is produced by the European Mathematical Society, the Fachinformationszentrum (FIZ) Karlsruhe, Germany and other Editorial Units all over Europe.

[1] http://www.ams.org/msc/

MSC covers all branches of pure and applied mathematics, including probability and statistics, numerical analysis and computing, mathematical physics and economics, systems theory and control, information and communication theory. MSC underwent in time a number of revisions; the latest version came valid in January 2000, so it is called MSC2000.

On the side of mathematics education, the **Zentralblatt für Didaktik der Mathematik Classification Scheme** [2] is used for the bibliographic database *MATHDI*, which is edited by the European Mathematical Society, FIZ Karlsruhe, and Zentrum für Didaktik der Mathematik at Karlsruhe Universit, in cooperation with Math Doc Cell (France)

1.2 – Schemes for computing, physics, control and information technology

In the field of computing, including hardware, software, networking, theory, methodologies and applications, the most important tool is the **Computing Classification System**.[3]
It is developed by the Association for Computing Machinery (USA) to classify items in the directories *Computing Reviews* and *Guide to Computing Literature*, which are edited by the same body.
Section 68 *Computer Science* of MSC was designed in rather tight matching with a great part of CCS.

In the fields of theoretical, experimental and applied physics and astronomy we have the **Physics and Astronomy Classification Scheme (PACS)**.[4] Section 02 *Mathematical methods in physics* of PACS closely resembles the top level codes for pure mathematics, probability and statistics of MSC. PACS is prepared and revised, at least biennially, by the American Institute of Physics.
A version of PACS is established as **Section A** of **INSPEC Classification**.[5] INSPEC is a bibliographic information service provided by the Institution of Electrical Engineers (UK). It covers physics, electrical engineering, electronics, communications, control engineering, computers and computing, and information technology.
INSPEC Classification has three other major sections:

- **Section B: Electrical & Electronic Engineering**
- **Section C: Computer & Control**
- **Section D: Information Technology**

1.3 – Schemes for economics

The fields of economics are increasingly involved in mathematical arguments, both in theoretical and specific topics; and conversely, mathematical problems and theories even more often arise from economic domains.

[2] http://www.mathematik.uni-osnabrueck.de/projects/zdm/
[3] http://www.acm.org/class/1998/
[4] http://www.aip.org/pubservs/pacs.html
[5] http://www.iee.org.uk/publish/inspec/docs/classif.html

This can be seen by the place mathematical topics take in the **Journal of Economic Literature Classification System**,[6] developed by the American Economics Association for its indexing journal and for the corresponding *EconLit* database.
Such topics are mostly located in the 62 *Statistics*, 90 *Operations research, mathematical programming*, and 91 *Game theory, economics, social and behavioral sciences* sections of MSC2000.

1.4 – Discipline specific and general schemes

Besides these, many other subject classification schemes exist for use in any scientific discipline or field of disciplines.
Yet other schemes are the general ones, not oriented to specific disciplines, such as **Dewey Decimal Classification**.[7]

2 – Classification schemes: from structure to browsing

2.1 – The common structure of subject classification schemes

The structure of subject classification schemes, be they discipline specific or general, is essentially the same: a relational system of *categories*, identified by alphanumerical *codes*, whose meaning is specified by *descriptions* or scope notes in some natural language (primarily, for current scientific research, English; translations and multilingual editions are frequently made available).
Generally there is one main relation, which in most cases is tree-shaped (monohierarchical, or, simply, hierarchical) and the categories are called *nodes*. Sometimes, however, the main relation is a more relaxed partial order, allowing nodes to be under more than one node (so the relation is called multihierarchical).
Other relations are considered as cross-references, allowing connections between diverging paths of the main relation-
Subject classification schemes vary in time through succeeding versions; one version keeps valid for indexing and searching in a bibliographic database for a more or less long period of years. Two subsequent versions can be related by linking categories in the older and the newer version which hold some correspondence in meaning, even if the relation may not be one-one, or structure preserving, due to splits, merges, reorganizations, deaths and births of topics, as represented in the positions of the two versions.

For example, **Mathematics Subject Classification** has 5531 categories in a three-level hierarchy. The top level counts 63 nodes. Cross-references, often equipped with explanatory

[6] http://www.aeaweb.org/journal/elclasjn.html
[7] http://www.oclc.org/dewey/products/index.htm

text ("For …") are of the following types: *see also* – *see mainly* – *see*. Some notes for coordinate indexing (and searching) are present.
Physics and Astronomy Classification Scheme has a four-level hierarchy. The top level counts 10 nodes, the second level 66 nodes.

2.2 – From structure to browsing

Due to their structural features, subject classifications are effective tools for browsing and searching in bibliographic databases, catalogs and other kinds of metadata repositories.

Moreover, subject classifications can set up knowledge organization tools for lexical collections extracted from metadata or fulltext databases, for terminologies, glossaries, dictionaries or encyclopedias, surveys, up to distributed libraries of natively digital documents or digitalized paper document. The set of descriptions of a classification scheme is itself a primary terminological resource.

2.3 – H-volumes in *The Scientific Classifications Page*

Different modes in browsing subject classifications can be exploited by hypertextual techniques.
We managed to produce various tools to demonstrate some of these modes.
The Scientific Classifications Page [8] collects such tools. It is presented both in English and in Italian language. It includes the following sections:

- *The Mathematics Classification Page*
- *Mathematics Subject Classification MSC and Dewey Decimal Classification DDC*
- *Relating Scientific Subject Classification*

The tools we produced consist of systems of syntactically simple but highly connected and coordinated HTML pages, called *h-volumes*.
H-volumes can amount even to thousands of files, written in plain HTML with simple JavaScript routines; in our working environment they are generated by a pool of standard C programs, starting from ASCII files, which present lists of records without redundancies and glossaries concerning attribute values.
H-volumes can be employed to display any kind of structured information set, such as directories, biographical collections, metadata collections, databases, glossaries, dictionaries, encyclopedias, etc.

The actual production of h-volumes starts from ASCII files obtained by manipulating existing data sets and texts, in particular available Web pages. This preparation activity is worked out partly by hand (i.e. using interactively some flexible source editor), partly making use of text processing procedures developed contextually to the development of procedures for HTML page generation.

[8] http://www.math.unipd.it/~biblio/math/eng.htm

Let's now turn to see the sections of *The Scientific Classifications Page* in some detail.

The Mathematics Classification Page

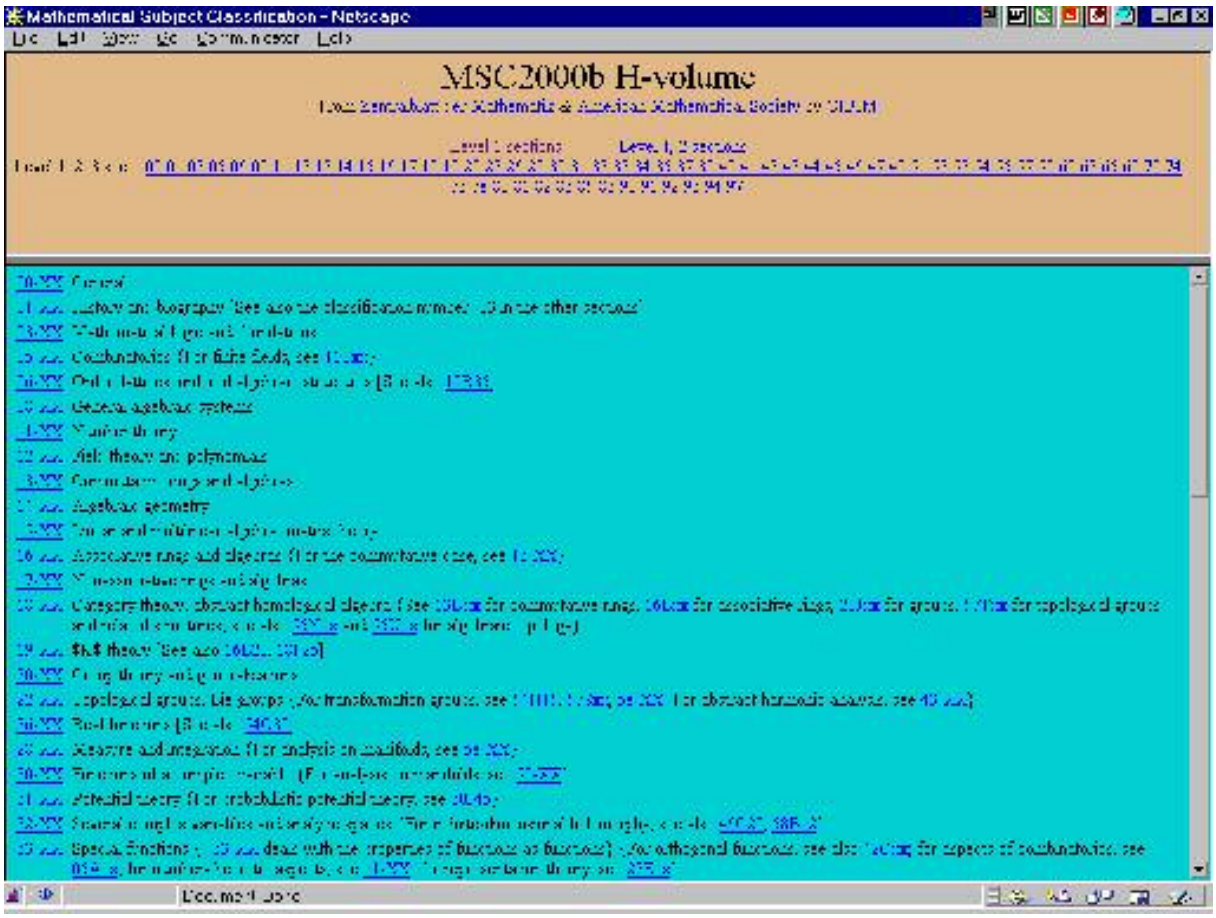
The Mathematics Classification Page collects six hypertextual frame presentations of the latest version of Mathematics Subject Classification, MSC2000.

From an ASCII file containing the whole MSC2000, a simple frame presentation was obtained. From the same file a double view presentation was obtained too.

The former process, generating a simple frame presentation, was worked out on a file containing an Italian translation of MSC2000, while, by processing the two files in combination, we obtained a simple frame presentation, which displays interleaved English and Italian texts.

From a file resulting from a comparison of MSC2000 with the 1991 version, we obtained a simple frame presentation which includes changes from MSC 1991:

Finally, from the combination of the first ASCII file with a file which contained data about subject-specific pages of relevant Websites, we obtained a simple frame presentation, with guide pages linking to those subject-specific pages of Websites.

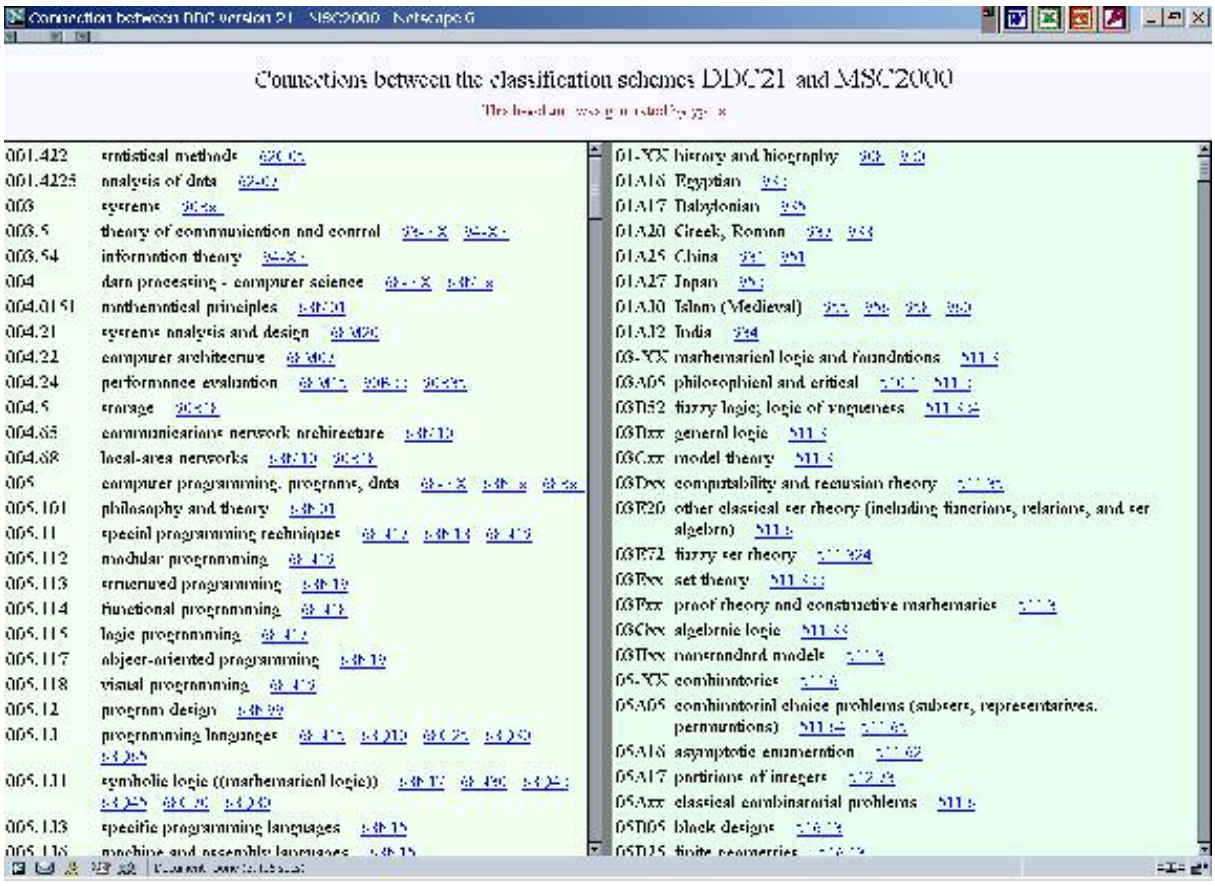


[Fig. 1]

This is an example of simple frame presentation.

The top frame is a sort of Table of Contents, which gives access to different slicings of the scheme: single list presentations of the classification categories at level 1 and 1-2, and an indexed set of list presentations which covers the whole scheme. For the latter, the top frame displays the list of the first 2 digits of the codes of the 63 level 1 categories; each item in the list points to a page which is displayed in the frame below, containing a list presentation of the subtree below the indicated level 1 category.
In this way, the long list of all the classification categories is divided into a number of sublists, so you can browse the classification scheme by transferring only files of moderate size.

On the other hand, double or multiple view presentations can be exploited to navigate through transversal links either inside one version of a classification scheme or among more schemes or versions: you can move to and from parallel views of them.



[Fig. 2]

Here is an example of double view presentation, showing connections between categories from the Dewey Decimal Classification, 21st edition, and MSC2000.

The double view presentation included in The Mathematics Classification Page is actually a duplicated simple frame presentation of the whole tree of MSC2000 in English which allows

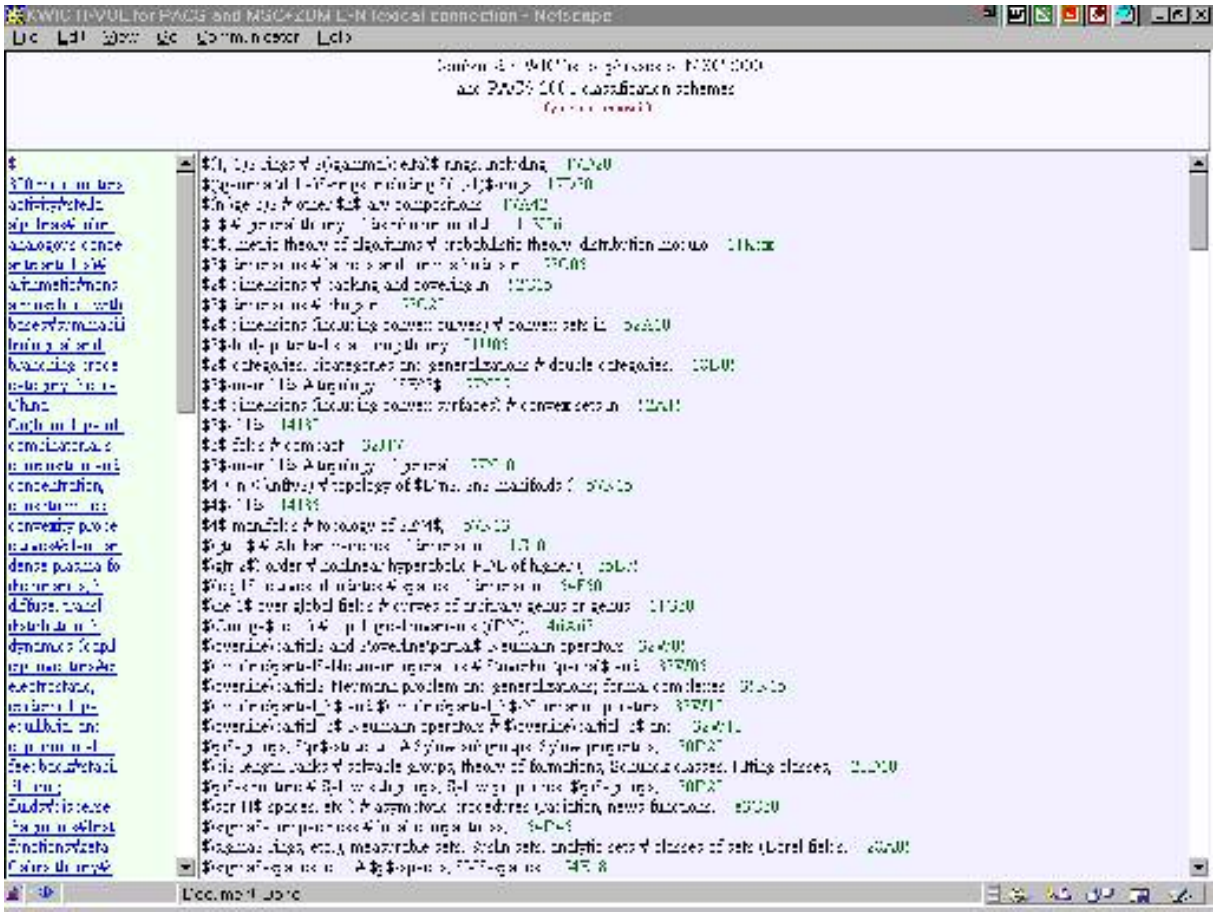
walking through cross references while keeping vision of the contexts of both endpoints of the selected cross references.

Mathematics Subject Classification MSC and Dewey Decimal Classification DDC

The *Mathematics Subject Classification MSC and Dewey Decimal Classification DDC* section of *The Scientific Classifications Page* includes two English language presentations:

- the just shown page of connections between categories from the Dewey Decimal Classification, 21st edition, and MSC2000
- a KWIC list h-volume for the combined set of descriptions of:
  - a revision proposal for the 510 DDC section, Mathematics, presented in January 2001
  - MSC2000

The sections E - N of the ZDM classification, encoded as 97E - 97N in the MSC style.



[Fig. 3]

KWIC list h-volumes (as in Fig. 3) are devised for discovering textual similarities among subject descriptions in one or more classification schemes or versions, in order to obtain suggestions about possible affinities of contents.

A KWIC list (KWIC shortens KeyWords In Context) presents every description through its circular permutations, beginning with a significant word or phrase; the overall list is ordered along the list of significant words.

By a method similar to that employed for simple frame presentation, long ordered list, as generally a KWIC list is, can be endowed with some sort of distributor allowing to reach quickly determined points or sections of the long sequence. A distributor can be built with pointers to initial letters, initial words of paged sections, sublists dealing with particular categories of entities. The list of permuted descriptions, subdivided into smaller manageable lists, is displayed on the right, while the distributor appears in the left frame.
KWIC lists may not be intended for the end user, rather as a help in establishing structured connections within or among classification schemes. This activity, although can greatly benefit from automated techniques, requires an amount of field specific knowledge which can't be automated, at least with the current technologies. The connections so discovered can be subsequently displayed through multiple view presentations of the involved classification schemes.

Relating Scientific Subject Classifications
The *Relating Scientific Subject Classifications* section of *The Scientific Classifications Page* contains a set of English language presentations (in one case bilingual):

- a double view presentation, showing connections between categories from the ACM Computing Classification System (1998), and MSC2000
- separate KWIC lists of descriptions of MSC2000, of PACS 2001, of ACM Computing Classification System (1998)
- combined KWIC list of descriptions of MSC2000 and PACS 2001, and of MSC2000 and ACM Computing Classification System (1998).

2.4 – Towards a presentation generating language

The h-volumes we produced are not intended to be taken as ultimate references, but as prototypes capable to clarify the real problems to face for the production of more complete and professional h-volumes and to test their effectiveness as documentation tools.
In fact, the development of such prototypes brought to the specification of parametrization mechanisms, data structures and processing modes which induced to define a programming language oriented to the manipulation of hypertextual presentations and to displays of mathematical structures.

The definition and the implementation of an experimental language called TAMP (Text Analysis Manipulation and Presentation) was actually started up.
TAMP is aimed to the analysis of text files of specified format (TeX, HTML, XML, etc.), the organization of specific knowledge bases endowed with links to other Internet resources and their presentation through HTML pages.
The language is implemented by means of a single C program, called YP, reading and generating only plain ASCII files. The first input file, characterized by the extension .ypg, is the source file of the program to execute. Many other specific files pointed out in the program are read and written.
Such files contain either data or sources of specific programs, dedicated to generate HTML files or other publishable files (e.g. TeX files), to prepare intermediate files, e.g., lists following defined orderings and collecting items provided by partial unordered files (in

particular files extracted from Web pages), or to control manipulations of some types (actually few) of mathematical structures starting from relatively simple expressions of basic ones in order to produce readable presentations of significant structures, possibly in a good consulting context.
The implementation is only at a “less than 1 version” and is poor in many respects, but has some peculiarities that allowed the production of practical Web pages and whose developments seem worthy of investigation.

The language can control many data types: the basic ones are integers (but not yet real numbers) and strings; it controls aggregates of basic data as sequences, tables and sequences of sequences. Moreover it's possible to manipulate some specific presentation structures (indexing KWIC lists, glossaries, etc.) and the representations of specific mathematical structures (permutations, partitions, graphs, trees, paths in combinatorial plane, etc.).
While a good choice of operators on basic data types and their aggregates is provided, only few operators acting on specific structures are implemented. On the other hand the implementing program YP has good extensibility features: the data types are parameterized, simple schemes allow the introduction of identifiers and general functional characteristics of new operators and their actions can be implemented in routines whose collocation and role are not difficult to tune with the characteristics of existing operators.
Among richer data types the language provides some kinds of constructors, composite entities targeted to build presentation structures. A typical example is given by the so called KWIC engine: its definition requires to specify the fields of a flat file, the catalogued routines charged to distinguish and accept these fields, the catalogued routines commissioned to build the different fields of final KWIC items and the parameters required by some routines. Specific statements allow to activate the constructors giving the possibility to choose for them parameters such as schemes controlling files to be generated and prefixes of their names.

An important characteristic of the language is the possibility to define automata at different levels of generality. The automata of the more general type can be defined by a specific rich jargon opening the possibility to determine effective models of acceptors, transducers, text analysers and text generators, typically through successive refinements.
Moreover, the translator of the proposed language can be used with a versatile preprocessor allowing substitutions, inclusions, selections and iterations of good reach: its control structures can act on variables concerning strings, integers and files. This preprocessor limits the actual major language drawback, i.e. lack of modularity. A group of statements that would be natural to encapsulate in a module can be recorded in a file endowed with dummy strings: this file can be included in other source files, either in the main one or in a file that can be included similarly.

3 – The OAI framework

3.1 – E-print communication: tools and networking architectures

Scientific research relies heavily on the rapid dissemination of results. So the slow formal process of submitting papers to journals has been augmented by other, more rapid, dissemination methods.
Originally dissemination involved printed documents, such as technical reports and informal conference papers.
Then researchers started taking advantage of the Internet, putting papers on ftp sites and later on various web sites. But these resources were fragmented. Searching through them resulted to be very difficult, and there was no guarantee that information would be archived at the end of a research project.
Different strategies for scientific research communication via e-prints were have been developed in time, which involve:

- small specialized archives
- centralized archives such as arXIv [9] for physics and related disciplines, mathematics, nonlinear sciences, computer science; and CogPrints [10] for cognitive science, artificial intelligence, computational linguistics and neurosciencesingle or networked institutional archives, such as NCSTRL [11] and the ERCIM Technical Reference Digital Library [12] for computer science and mathematics distributed networks connected by some interoperability protocol, such as RePEc [13] for economics, and DoIS [14] for library and information science
- umbrella servers, such as MPRESS [15] for mathematics
- servers connected to groups of journals or sponsored by commercial publishers, etc.

Web search and cash engines like Researchindex (formerly Citeseer),[16] provide a solution which has been appreciated especially by people in the computing area. E-prints posted in personal homepages without any specific care about metadata are harvested and cashed; the service is comprehensive with reference linking.

3.2 – The Open Archives Initiative

The Open Archives Initiative (OAI) [17] is an international effort to develop interoperability standards for disseminating content over the Web. OAI stresses the separation of being a *data*

[9] http://arXiv.org
[10] http://cogprints.soton.ac.uk
[11] http://www.ncstrl.org
[12] http://www.iei.pi.cnr.it/DELOS/EDL/ETRDL_Con/
[13] http://www.repec.org
[14] http://dois.mimas.ac.uk/
[15] http://mathnet.preprints.org
[16] http://citeseer.nj.nec.com
[17] http://www.openarchives.org

provider (i.e., publisher) and being a *service provider* (i.e., interface for search, browsing, reference linking). On the other hand, nothing prevents the same system to embody and integrate both functions. It is even possible for individual researchers to develop personal open archives, which can be accessed to build tailored personal web sites and other services, as well as harvested into department archives.

The base concept of the OAI is metadata harvesting, which is realized in the *OAI Protocol for Metadata Harvesting*. [18] So it no longer matters *where* papers are archived; the papers in all registered OAI-compliant archives can be harvested using the OAI protocol into one global "virtual archive" by Open Archives service providers.

3.3 – OAI compatible refereed self-archives: the EPrints 2 software

EPrints [19] is a free (General Public License) software for managing e-prints archives, developed at the Electronics and Computer Science Department of the University of Southampton (UK).
It is aimed at organizations and communities rather than individuals. It provides an interface for system administrators, for archive editors to process submissions, for authors to deposit papers, and for users to access papers by searching or browsing metadata.
The system comes configured to run an institutional pre-prints archive, but can be reconfigured with utterly different metadata fields and content.
Any version of EPrints is fully interoperable with the current *OAI Protocol for Metadata Harvesting*.

4 – Conclusions

Our work has been directed to the definition of text processing methodologies for the development of hypertextual presentations of complex documentation structures. Such presentation modalities can enrich the browsing functionalities of archives and service providers in the OAI framework, allowing a full network of bridges among specific subject areas to guide advanced research communication activities.

In particular, we are investigating the possibility of providing the EPrints software with tools modeled on the experimental ones we produced for the *Scientific Classification Page*.
Centering with Mathematics Subject Classification, bridges can be launched and passed through inside mathematics and among the disciplines that live and develop with mathematics. This is equivalent to say that bridges can be launched all over the world of scientific and technological knowledge, if we are aware of the dynamics that mathematical disciplines are ever more moving in modeling and computing activities for every field of human knowledge.

[18] http://www.openarchives.org/OAI/openarchivesprotocol.htm
[19] http://www.eprints.org

References

Antonella De Robbio, Dario Maguolo, Alberto Marini
Scientific and General Subject Classifications in the Digital World
High Energy Physics Libraries Webzine, Issue 5, November 2001
http://doc.cern.ch/heplw/5/papers/4/

Alberto Marini
Text Processing for Presentation and Manipulation of Mathematical Resources
Paper presented at the Workshop "Electronic Media in Mathematics", Coimbra (Portugal), September 13-15 2001
http://www.mat.uc.pt/EMM/index.html