HAL
open science

Digital Library and Document Server

Abdel Belaïd

To cite this version:

Abdel Belaïd. Digital Library and Document Server. International Workshop on Technology Development in Indian Language, ISI, Calcutta, 2001, Calcutta, India, 39 p. inria-00107519

HAL Id: inria-00107519

https://inria.hal.science/inria-00107519

Submitted on 19 Oct 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Digital Library and Docum
Server

Calliope, Muse of eloquence and poe

The Calliope Project

A. Belaïd

The Framework

R&D Collaboration

- IMAG + INRIA + RXRC (Xerox)
- First collaboration success : Callimaque
  - 3000 thesis & technical reports : 1950 - 1993
  - RXRC : multilingual interrogation for anglophone readers
  - IMAG : 320 journals, annual budget 150 000 $

The problem

- INRIA Displacement, deprived researchers of doc access

The Solution

- Not duplicate the center
- Automatic system : access to IMAG document center

Constraints

nfrastructure :

- Investment : to minimize the cost in material
  - Use : Net infrastructure (connection flow 128 K bits/s)
  - Permit : heterogeneous client station (Macintosh, PC, Unix)
- Quality : to deliver articles with a quality similar to the photocopy
  - Postscript level 2

esearch : to take advantage of this opportunity

- To go beyond the paper model
  - To better apprehend the mass of scientific information
  - To integrate sources of heterogeneous documents
  - To open the research on data indexing, data mining, ...

The Calliope service

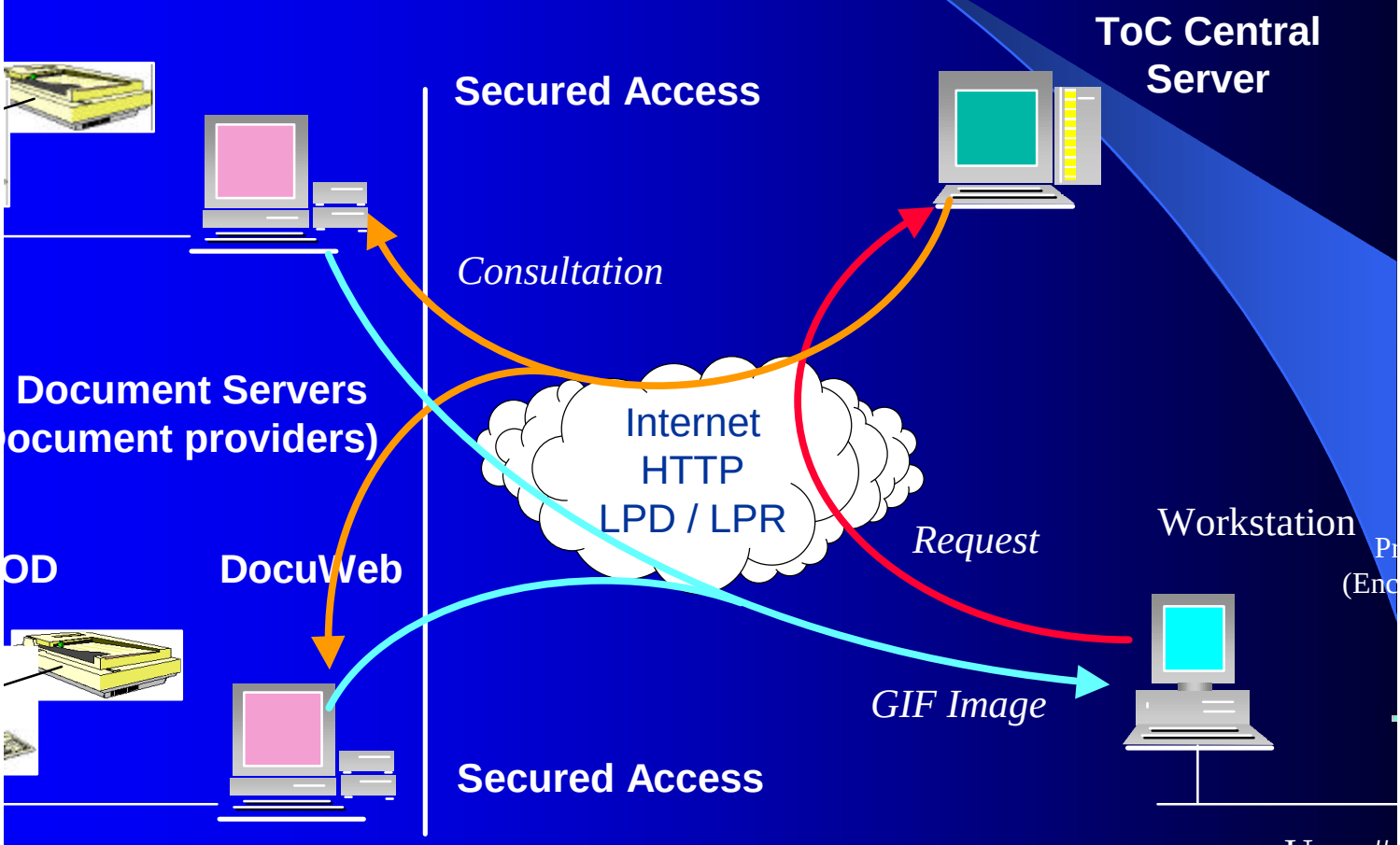
The solution

- To develop the concept of telephotocopy

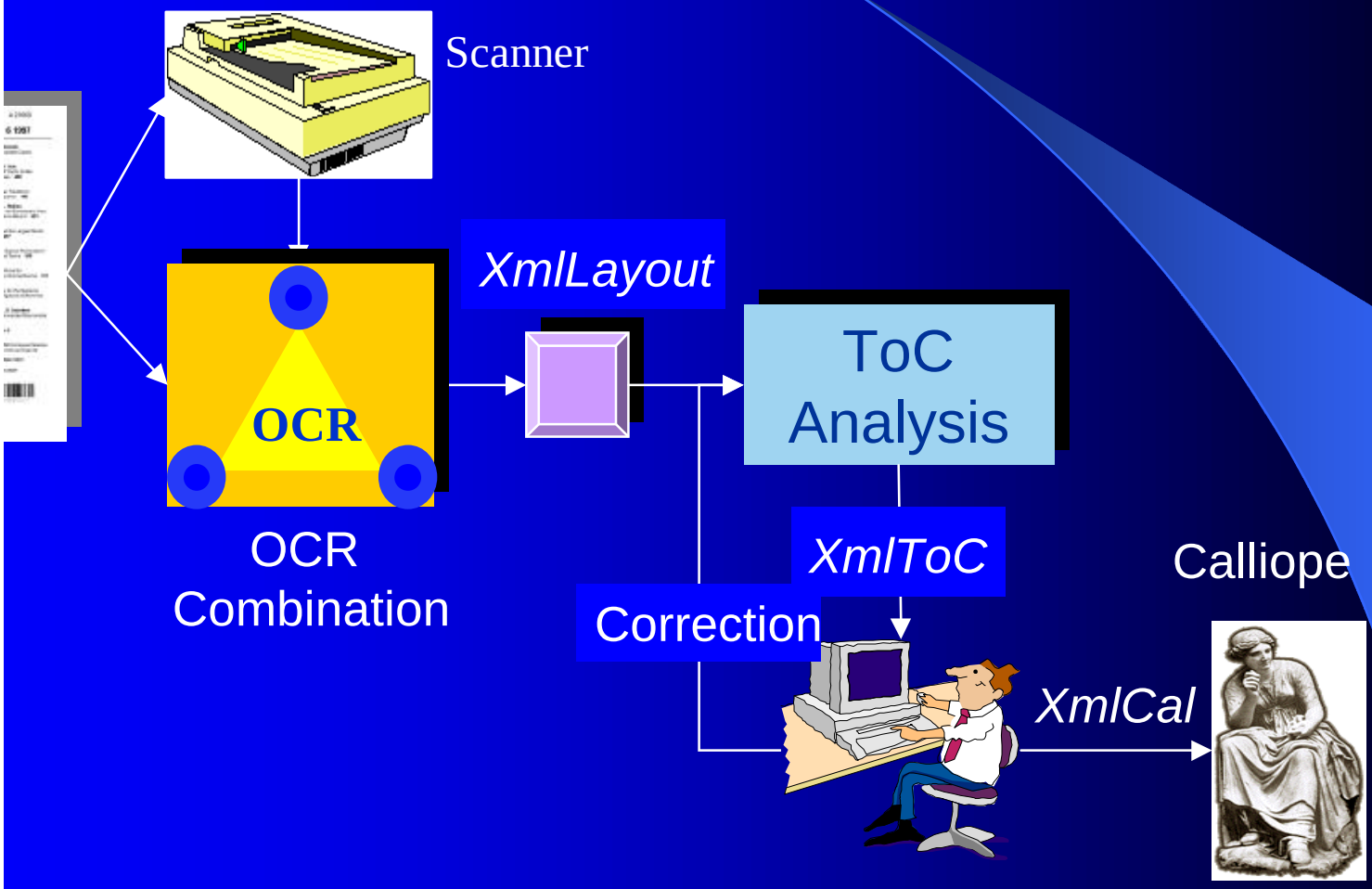
Functionality

- Access to table of contents server
- Asking of telephotocopy by :
  - Simple click on the article
  - Textual Retrieval on paper titles, authors, journal title
- Services
  - Availability signaled by email
  - Get the paper in an image format
  - Printing at the request of part or total paper
  - Electronic subscription

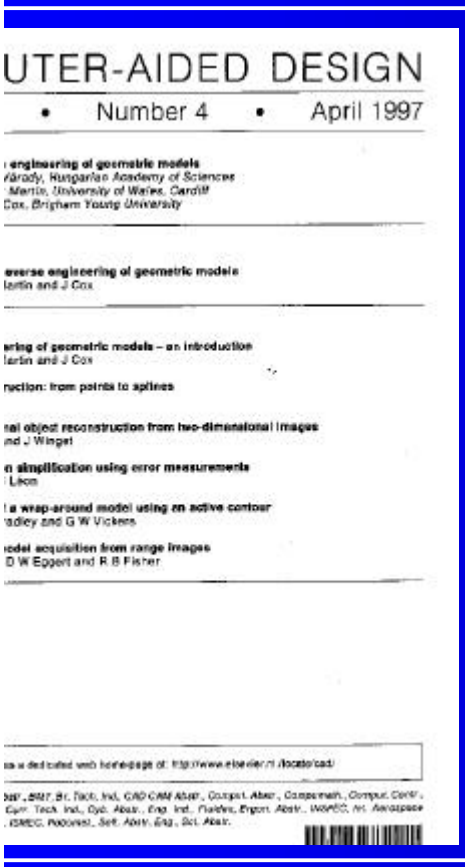
Architecture



Recognition



ToC Examples


UTER-AIDED DESIGN

• Number 4 • April 1997

engineering of geometric models
Ároody, Hungarian Academy of Sciences
Martin, University of Wales, Cardiff
Cox, Brigham Young University

verse engineering of geometric models
Martin and J Cox

ering of geometric models – an introduction
Martin and J Cox

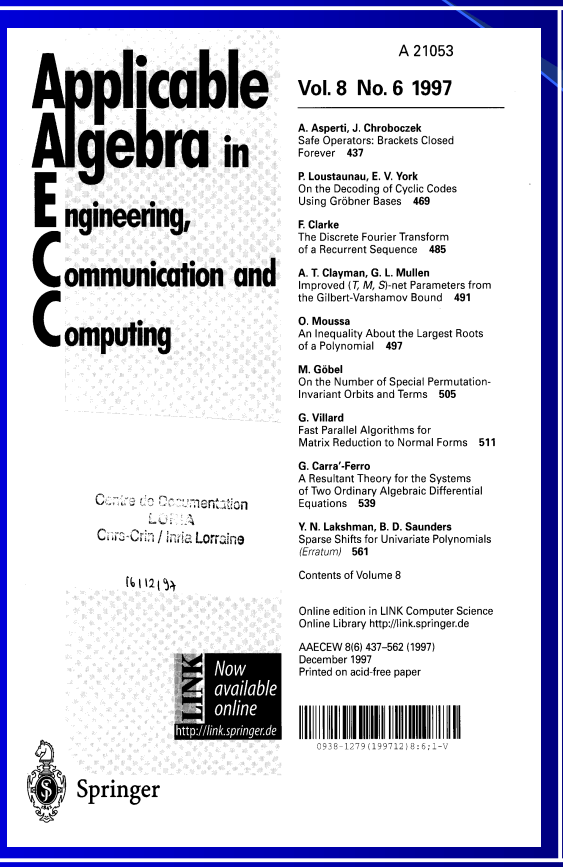
nction: from points to splines

al object reconstruction from two-dimensional images
nd J Winget

n simplification using error measurements
Leon

a wrap-around model using an active contour
adley and G W Vickers

odel acquisition from range images
D W Eggert and R B Fisher


Applicable Algebra in Engineering, Communication and Computing

A 21053

Vol. 8 No. 6 1997

A. Asperti, J. Chroboczek
Safe Operators: Brackets Closed Forever 437

P. Loustaunau, E. V. York
On the Decoding of Cyclic Codes Using Gröbner Bases 469

F. Clarke
The Discrete Fourier Transform of a Recurrent Sequence 485

A. T. Clayman, G. L. Mullen
Improved (*T, M, S*)-net Parameters from the Gilbert-Varshamov Bound 491

O. Moussa
An Inequality About the Largest Roots of a Polynomial 497

M. Göbel
On the Number of Special Permutation-Invariant Orbits and Terms 505

G. Villard
Fast Parallel Algorithms for Matrix Reduction to Normal Forms 511

G. Carrà-Ferro
A Resultant Theory for the Systems of Two Ordinary Algebraic Differential Equations 539

Y. N. Lakshman, B. D. Saunders
Sparse Shifts for Univariate Polynomials
(Erratum) 561

Contents of Volume 8

Online edition in LINK Computer Science Online Library http://link.springer.de

AAECEW 8(6) 437–562 (1997)
December 1997
Printed on acid-free paper

0938-1279(199712)8:6;1-V

Centre de Documentation
LORIA
Cirs-Crin / Inria Lorraine

Now available online
http://link.springer.de

Springer

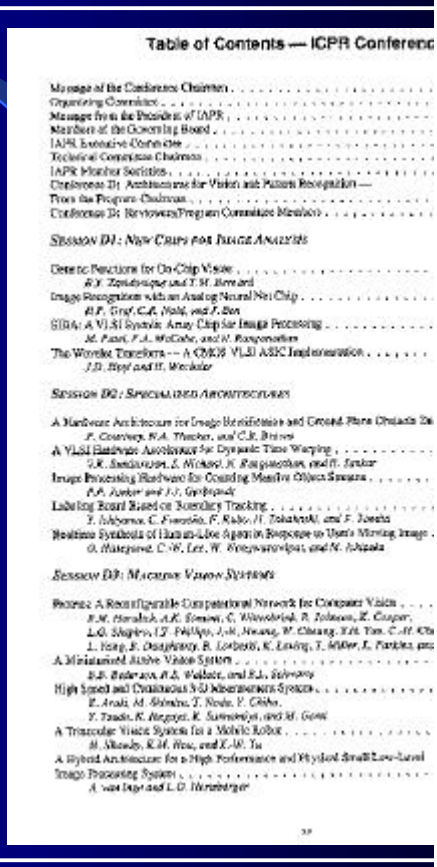


Table of Contents — ICPR Conferenc

Message of the Conference Chairman
Organizing Committee
Message from the President of IAPR
Members of the Governing Board
IAPR Executive Committee
Technical Committee Chairman
IAPR Member Societies
Conference D: Architecture for Vision and Pattern Recognition —
From the Program Chairman
Conference D: Reviewers/Program Committee Members

Session D1: New Chips for Image Analysis

Genetic Functions for On-Chip Vision
Image Recognition with an Analog Neural Net Chip
SIMA: A VLSI Systolic Array Chip for Image Processing
The Wavelet Transform — A CMOS VLSI ASIC Implementation

Session D2: Specialized Architectures

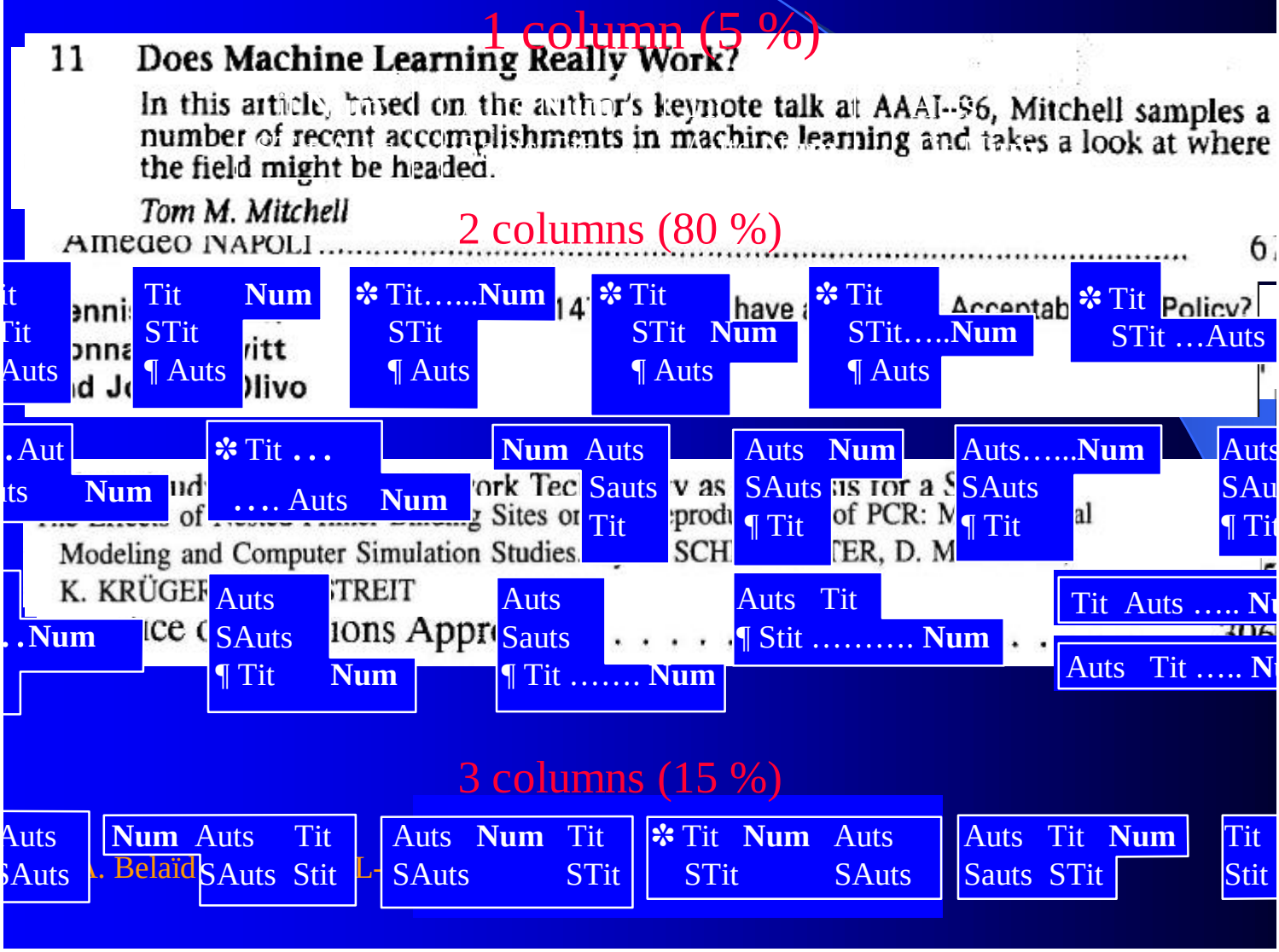
A Hardware Architecture for Image Identification and Ground Plane Obstacle De
A VLSI Hardware Architecture for Dynamic Time Warping
Image Processing Hardware for Counting Massive Object Streams
Labeling Board Based on Boundary Tracking
Realtime Synthesis of Human-Like Agents in Response to User's Moving Image

Session D3: Machine Vision Systems

Pegasus: A Reconfigurable Computational Network for Computer Vision
A Minimanned Active Vision System
High Speed and Continuous 3-D Measurement System
A Trinocular Vision System for a Mobile Robot
A Hybrid Architecture for a High Performance and Physical Small Low-Level Image Processing System

Reference Structure

Top down approach

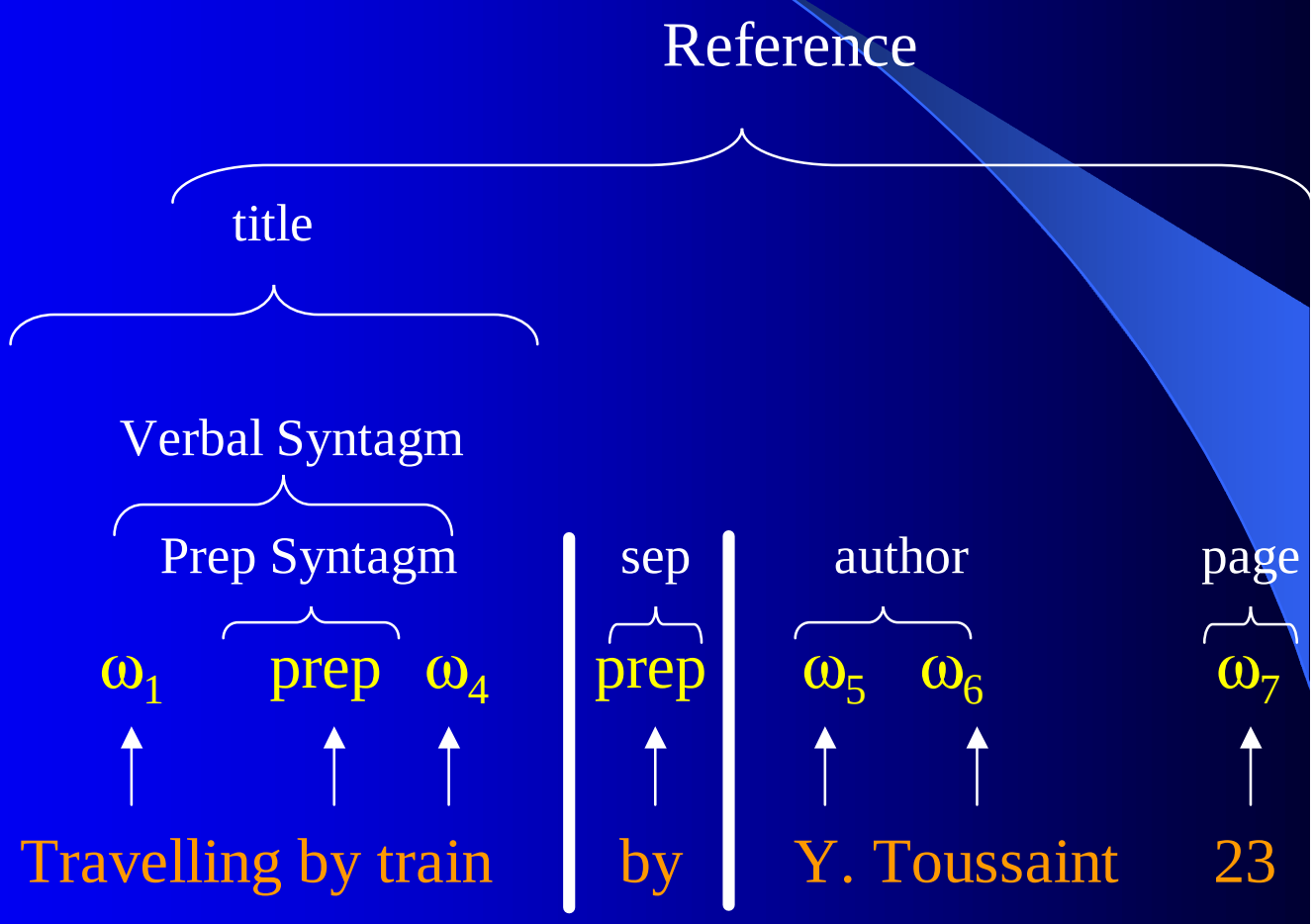


Idea : Bottom-Up Approach

Reference

Title	Author	Page
CW CW CW CW	IT PN	Num
A new tagging approach	A. Belaïd	23

Linguistic Help



Morpho-syntactical Labeling

Role :

- Annotate text to mark important elements of a corpus
- Associate these elements in more complex structures (terms) by syntactical agglutination

Initial application domain : terminology

- Search for terms in a document (comp of several words)

development/CN of/PREP systems/PCN

We can have :

Rule 1 : term ::= CN of/Prep PCN

Advantages

raining :

- Most of the systems work with a usable learning
- It is possible to enrich the system and to adapt it for its own use by doing an adequate learning

rediction :

- To have a system able to predict a label for an unknown term
- Example : Brill rule for French : **radical+ion Þ CN**

ocation Principle : labeling error correction by local rul

- **beginning of a sentence + capital Þ CN,**
- **capital in the middle of a sentence Þ PN**
- **JJ of/Prep CN Þ JJ = CN**

Application

l Representation for ation in Animats	J. Prescon
oring Strategies for dded Agents: Experiments nalysis	S. Ackin and R. Cohen

Prep p UN	UN . UN
CN Prep CN : PCN N	UN . UN CC UN. PN

Remarks

- Linguistic defects in lab. come from particularitie
  - Terms of the structure : not conform to text notion
  - Word structure : not conform to the morphology of th current language

S/UN ./. Ackin/UN

- Solutions :
  - Retrain the system: costly and not adapted
  - Choose a less fine grain for labeling and a set of adapted labels to the application

Adapted Labels

Physical Labels

Blank Line	BL	Beginning Line
Long Space	EL	Ending Line
Current Line	NL	Next Line
Numerical Line	PL	Previous Line
Dotted Line	PT	Punctuation
Tabulation	PU	Item

Morphological Labels

Common Noun	AR	Article
Proper Noun	FN	First name
Initial	CC	Connector
Preposition	UN	Unknown

Reference Location

Typographical aspect combination

ge number detection

Regularity

- Location : position, before/after TB/SL
- Alignment
- Continuity

eference beginning/ending determination

CL = BL : NS & (PL=SL | PL ' EL)

CL = EL : NS & (NL=SL | NL ' BL)

olumn identification

COL = {**CL** : ≅alignment, ≅position, >**TB**, <**TB**}

July 1992 • Vol.

ARTICLES

5 **Guest Editors' Introduction: Document Image Analysis Systems**

Lawrence O'Gorman and Rangachar Kasturi

10 **A Prototype Document Image Analysis Syste Technical Journals**

George Nagy, Sharad Seth, and Mahesh Viswanathan

Intelligent document segmentation cari bring electronic browsing within the reach c users. The authors show how this is achieved through document processing, analysis parsing the graphie sentence.

25 **An Interpretation System for Land Register Maps**

Luca Boatto, Vincenzo Consorti, Monica Del Buono, Silvana Di Zenzo,, Vincenzo Eramo, Alessandra Esposito, Francesco Melarne, Marco Meucci, Andrea Morelli, Marco Mosciatti, and Marco Tucci

The semantics of land register maps drive this document conversion system. Howev its methods of image representation, vectorization, and symbol recognition cari be g other classes of line drawings.

34 **Postal Address Block Location in Real Time**

Paul W. Palumbo, Sargur N. Srihari, Jung Soh, Ramalingam Sridhar, and Victor Dem

A postal automation system locates destination address blocks on letter mail pieces success rate. Pipelining and multiprocessor techniques achieve real-time processing

46 **Celesstin: CAD Conversion of Mechanical Drawi**

Patrick Vaxivière and Karl Tombre

A prototype CAD conversion system extracts higher level structures for knowledge based analysis. It recognizes such entities as screws, ball bearings, and shafts.

Primary Labeling

			CN SL SL CN SL
presentation for in Animats	Tony J. Prescon	85	AJ CN PR TB FN IT PN TB NS CN PR UN SL
Strategies for Agents: Experiments sis	Marc S. Ackin and l 'aul R. Cohen	125	CN CN PR TB FN IT UN CC UN CN CN PT CN CC CN SL
the competitors	Luc Steels	173	CN AR CN TB FN PN TB NS SL UN SL
Adaptations and ry Epistemology	James H. Fetzer	201	CN CN CC TB PN IT UN TB NS CN CN

Combination of linguistic aspects

orming rules :

- **Author(Aut): {IT ; FN ; PN}**
- **Title(Tit): {CW}-{negative cuttings} :DET,P**

egrouping rules :

- **Author : {Aut; CC ; BY ; PT}**
- **Title : {Tit ; CC ; PT}**

ontinuity rules :

- **Author : {IT ; FN} + EL + PN Þ Aut + EL + A**

ative Method for Improved Protein Structural Motif Recognition.
BERGER and M. SINGE

cal Adaptations and Lee K. Wang and James H.
onary Epistemology Fetzer

Author Correction :

Re-establishment of the syntax

glutination rules:

ut + UN + Aut Þ Aut

ension rules:

ut + UN + {PN} + DL|EL Þ Aut + DL|EL

icipation rules:

Tit + EL + TB + UN^{+} Þ Tit + EL + TB + $Aut^{Þ}$

$Aut^{Þ}$ + CC + FN|IT + UN Þ Aut

Bailey, 1. J. Hall, 75 CBL in engineering: students' us
S. Reed ancl C. J. Colbourn alearning resource on phase dia

ic Functions for On-Cli Vision.....
BY. Zavidavique and T.M. Bernard

Title Correction :

e-establishment of the syntag

glutination rules:

t^{+} + UN + Tit^{+} Þ Tit^{+}

t^{+}| $Tit^{Þ}$ + Prep + Aut + Tit^{+} Þ Tit^{+}

$t^{Þ}$ + Prep + Tit^{+} Þ Tit^{+}

ension rules + anticipation :

t + Prep|Det + UN^{+} Þ $Tit^{Þ}$

+ Prep|Det + Tit Þ Ü Tit

t + EL + TB + UN^{+} Þ Aut + EL + TB + $Tit^{Þ}$

ing and evolulion in networks

es extracted from the article of /IN Claude Allègre « What I want »

li /NN error /NN measure /NN approach /NN to /TO train /VB feedforwa

neural /NNP networiks /NNP for /IN classification /NN problems /NNS

nopapproach /NNP usingspreadsheetsfortheteaching /NNP of /IN statis

probability /VBP

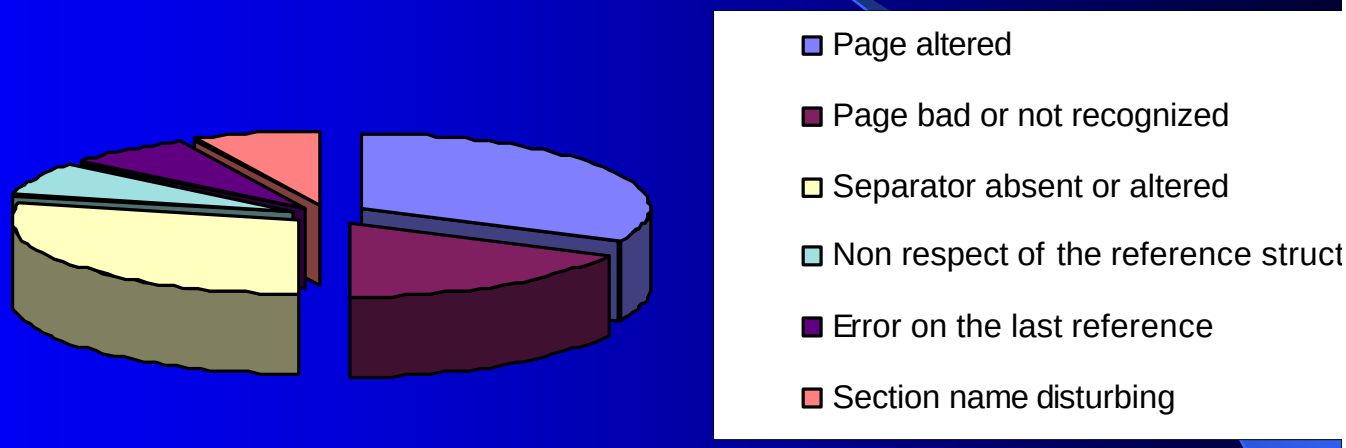
Model Generation

on	7	Tit PP
URGER s d'interaction une approche modulaire	11	Aut TB Tit
LLIETTAZ ux actionnels dans les interactions verbales nition de la dimension référentielle du discours	47	Aut TB Tit TB Stit PP
OBET tuation prosodique dans les dimensions périodique nationnelle du discours	83	Aut TB Tit TB Stit PP
MICHE sation polyphonique d'un fragment de débat onique	125	Aut TB Tit TB Stit PP
ULET sation polyphonique et l'organisation inférentielle ogue romanesque	149	Aut TB Tit TB Stit PP

Results (one OCR)

ocument Class	Author Field Number	Identified Authors	Article Number	Identifie Articles
Journal	1037	857	1256	1233
Journal	302	290	326	326
ceedings	204	152	207	194
ceedings	305	227	317	286
Magazine	159	151	168	159
al	2020	1677	2277	2198
		83%		96.5%

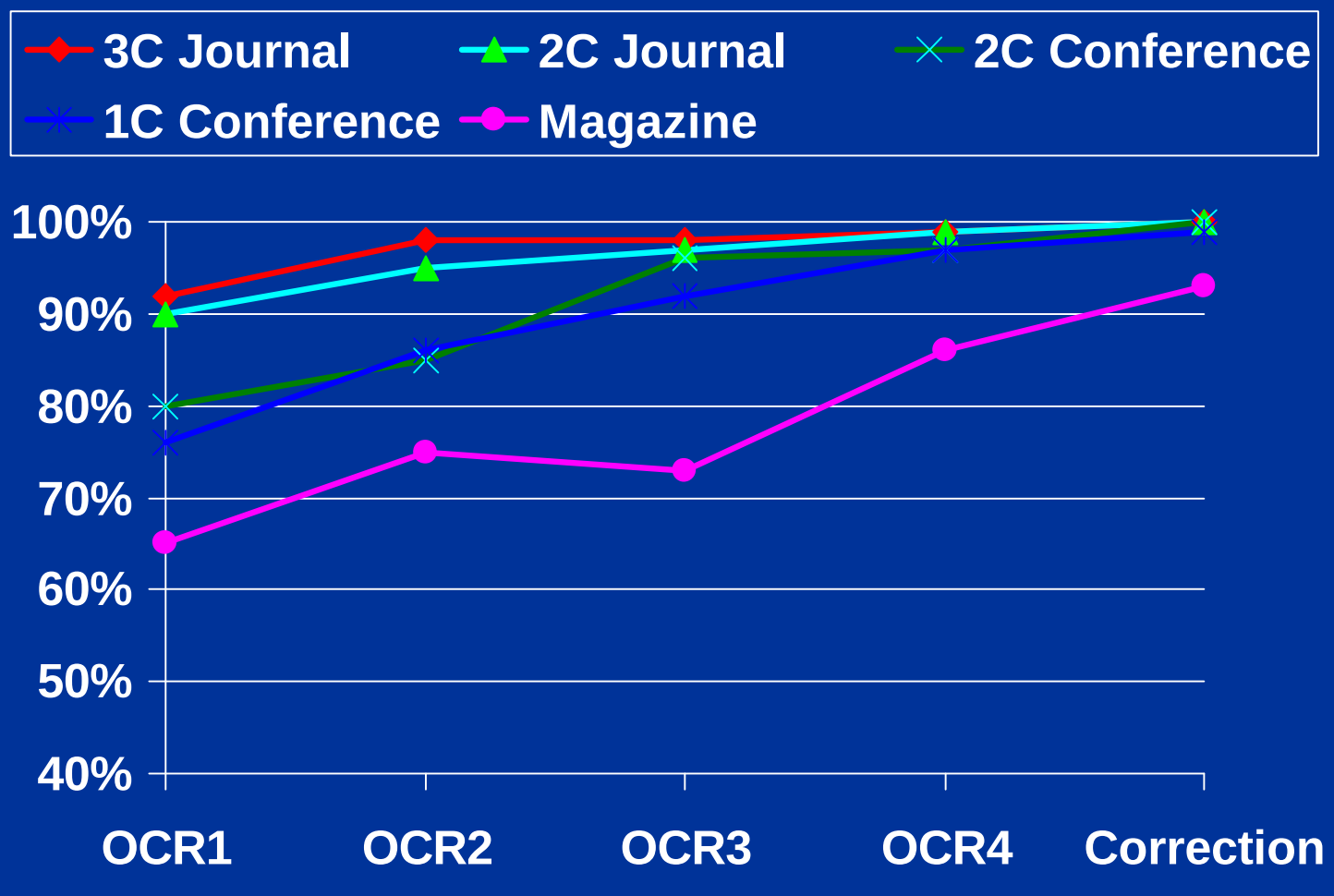
Defects



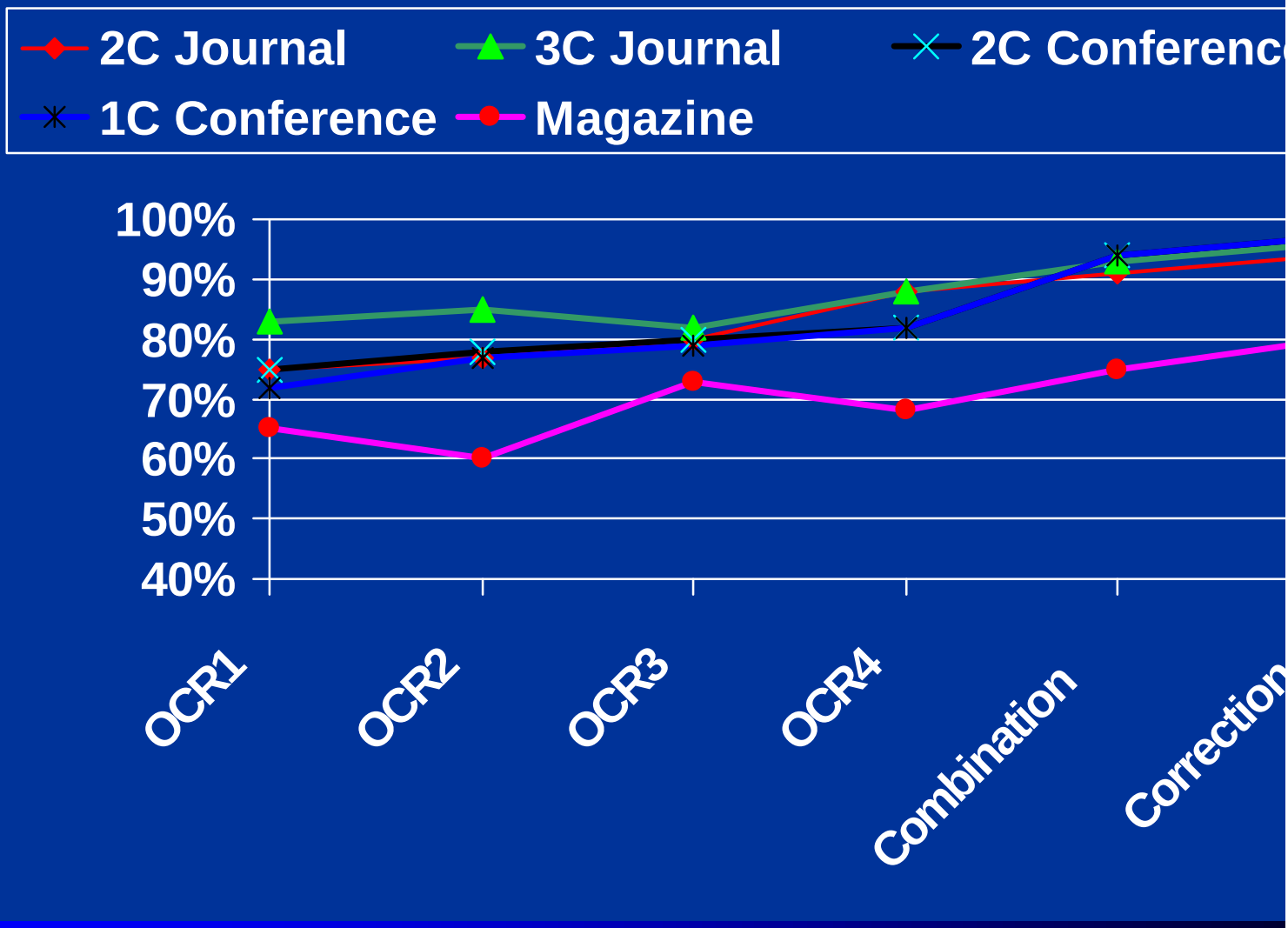
ion DS: Parallel Implementations
nces In Massively Parallel Computing :
view 95
.C. Sanz

ributing Authors.....95
uctions to Authors.....97

Segmentation



Structure


2C Journal
3C Journal
2C Conferenc
1C Conference
Magazine
100%
90%
80%
70%
60%
50%
40%
OCR1
OCR2
OCR3
OCR4
Combination
Correction

Conclusion and perspective

Prototypes :

- **INRIA Unit Library connection**
- **Tunisian University Library connection**

Research : Problem Serialization

- Content :
  - OCR Result Improvement
- Linguistics
  - Revise the tag names, more fine
  - Study of other linguistic extension rules

Scientometrics

How to measure the impact of an institution in one domain ?

Bibliometrics

Surface Info : distributions, frequencies per date, journal, author, domain

Lab Production, Country

Infometrics

Info on the content : to highlight themes, co-citations

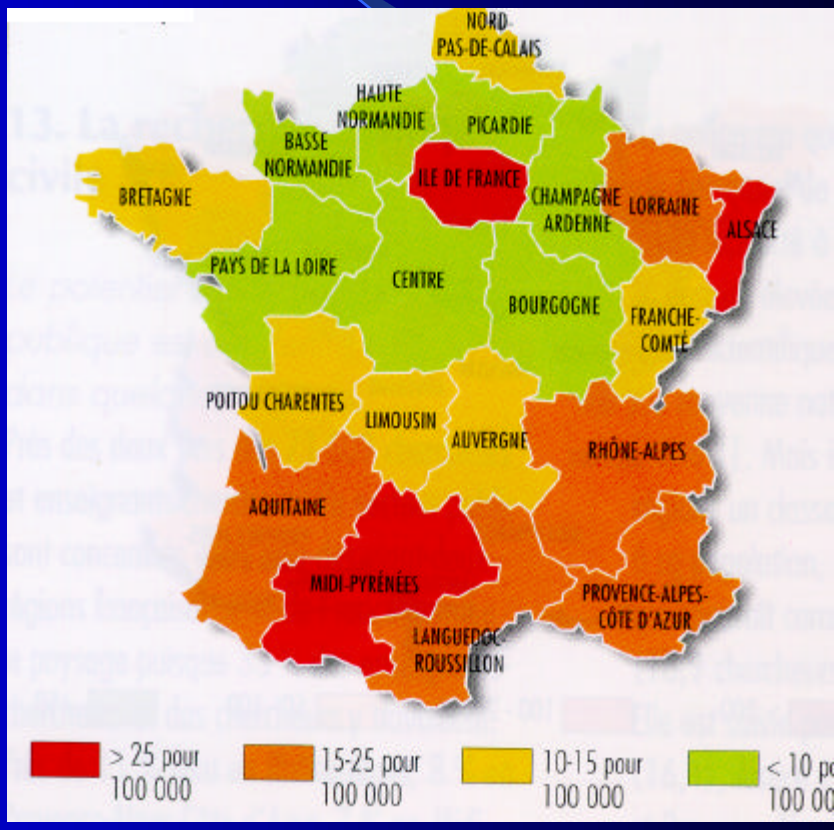
Relational Indicators between themes and authors

Bibliometrics

Nb of thesis / population

Predominance of the Ile le France

Effects on the neighbor regions



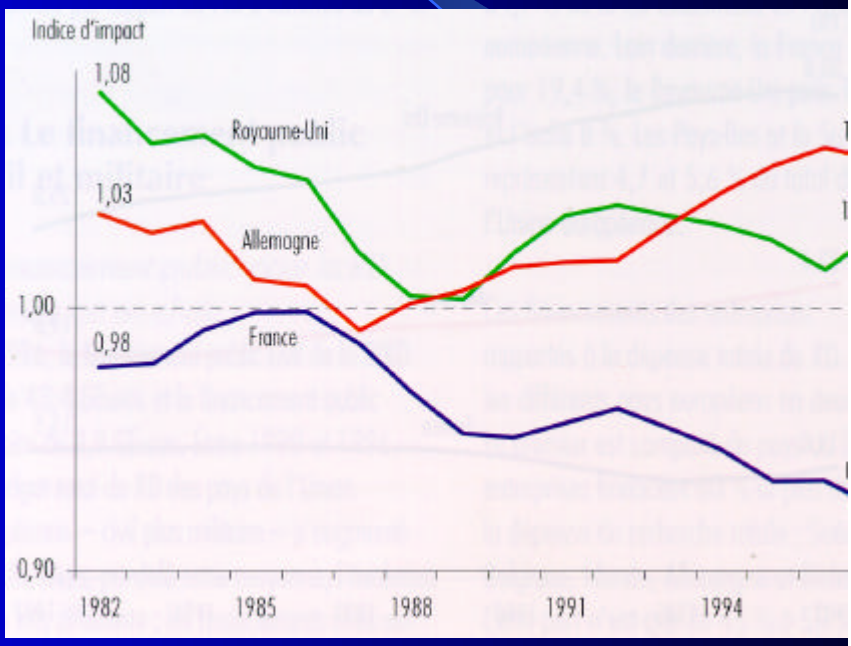
Bibliometrics for article referencing

mpact Index of the scientific ublications in the E.U.

mpact = average nb of references er article

)ne remarks a continuous ecreasing in France, i.e. french rticles are less and less read and eferenced

everal possible causes...



:rying to transform not only
size and position, but also some
using more than one harmonic.
r harmonics can better describe
s of the characters and would
nces to distinguish between
ers having close shapes.

& M Biquard, «Signaux
èmes Linéaires et Bruit en
ronique », Paris, 1992.

M, Lecolinet E, Cheriet M,
. C.Y, Automatic reading of
ve scripts using a reading model
perceptuel concepts, IJDAR,
me1, Number1, 1998, pp 3 - 17.

M.Hamanaka, K.Yamada,
ulumo, « On-line Japanese
acter Recognition Experiments
an Off-line Methid Based on
nalization-cooperated Feauture
action », ICDAR, IEEE, 1993,
04-207.

Authorrs (4) : Côte M., Lecolinet E., Cheriet M., Suen C. Y.
Title : Automatic reading of cursive scripts using a reading model and perceptuel concepts
Editor :

Month : Year : 1998 → Historical
Volume : 1
Number : 1
Publisher :
Address :
Pages : 3-17
Organization :
Booktitle : IJDAR → Indice d'impact
Series :
Note :

Occurrence + Co-occurrence

The EU Libraries Program

Telematics program : 1990-1994

- **Availability & accessibility** of modern library services
- More rapid penetration of I&CT in a **cost-effective way**
- **Standardization** required for resource sharing
- **Harmonization** and convergence of national policies

28 projects : 28 MECU

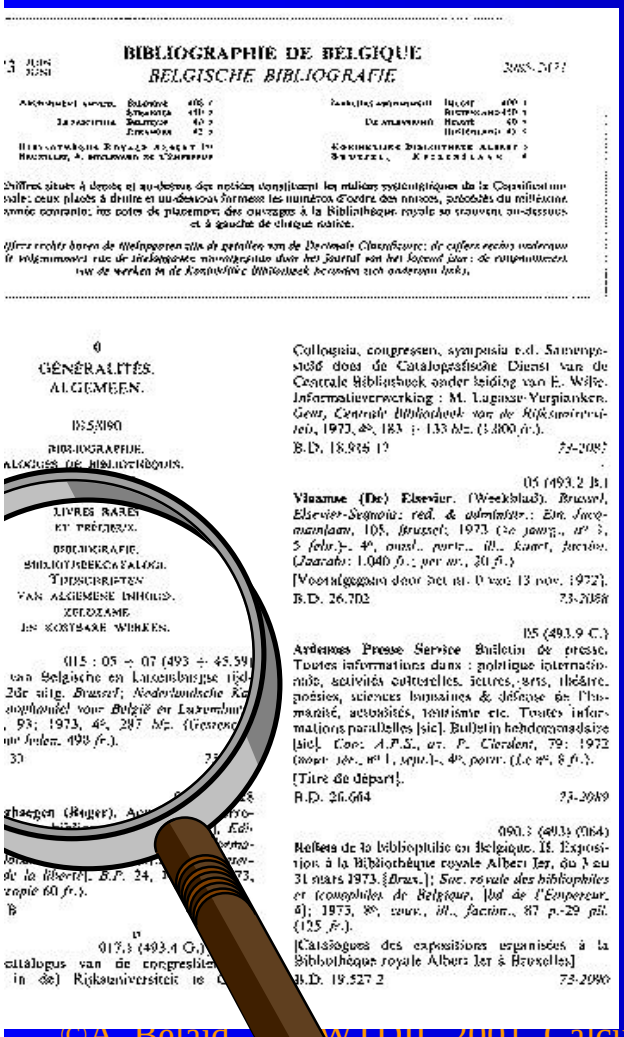
4 Action Lines

- To create, enhance and harmonize machine-readable bibliographies and union catalogues, tools for retrospective conversion of catalogues of important collections
- To impel the international library networking, providin access to source data
- Innovative services for libraries, using new technology
- Development and production of commercially viable product, services and tools specially designed for librar

Three Projects: OCR/ICR

- **BIBLIOTHECA**: Spain + Italy + France
  - Study of a pivot format (XML) for the representation of diffe records
  - Study of a system for indexing and classifying different documents
- **FACIT**: Denmark+Greece+Italy
  - Search for adapted OCR packages for retroconversion with la set of characters, and tools for fast and cheap mass conversion catalogues
- **MORE**: Belgium+France
  - Study of the role & use of dictionaries in the structure modelin and Recognition of catalogues by OCR techniques

Catalogue example



BIBLIOGRAPHIE DE BELGIQUE
BELGISCHE BIBLIOGRAFIE

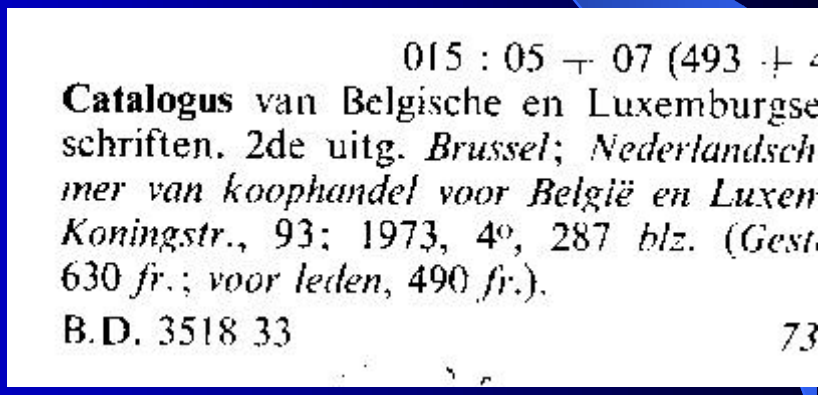
0
GÉNÉRALITÉS.
ALGEMEEN.

Collognia, congressen, symposia e.d. Samengesteld door de Catalografische Dienst van de Centrale Bibliotheek onder leiding van E. Wilte. Informatieverwerking : M. Lagasse-Vergiaanken. *Gent, Centrale Bibliotheek van de Rijksuniversiteit*, 1973, 4°, 183 p. : 133 blz. (3.800 *fr.*).
B.D. 18.936 12 73-2085

05 (493.2 B.)
Vlaamse (De) Elsevier. (Weekblad). *Brussel, Elsevier-Sequoia: red. & administr.: Em. Jacobsmaanlaan*, 105, *Brussel*; 1973 (*ste jaarg., n° 5, 5 febr.*)-, 4°, *ouvl., portr., ill., kaart, facsim.* (*Jaarabo*: 1.040 *fr.*; *per nr.*, 20 *fr.*)
[Voorafgegaan door Sel ni. 0 van 13 nov. 1972].
B.D. 26.702 73-2088

05 (493.9 C.)
Ardennes Presse Service Bulletin de presse. Toutes informations dans : politique internationale, activités culturelles, lettres, arts, théâtre, poésies, sciences humaines & défense de l'humanité, actualités, tourisme etc. Toutes informations parallèles [sic]. Bul)etin hebdomadaire [sic]. *Con: A.P.S., av. P. Clerdent*, 79: 1972 (*nouv. sér., n° 1, sept.*)-, 4°, *porte.* (*Le n°, 8 fr.*).
[Titre de départ].
B.D. 26.664 73-2089

090.3 (493) (084)
Reflets de la bibliophilie en Belgique. II. Exposition à la Bibliothèque royale Albert Ier, du 3 au 31 mars 1973. [*Brux.*]; *Soc. royale des bibliophiles et iconophiles de Belgique*, [*bd de l'Empereur*, 4]; 1973, 8°, *couv., ill., facsim.*, 87 *p.*-29 *pll.* (125 *fr.*).
[Catalogues des expositions organisées à la Bibliothèque royale Albert Ier à Bruxelles]
B.D. 19.527 2 73-2090



015 : 05 + 07 (493 + 4
Catalogus van Belgische en Luxemburgse
schriften. 2de uitg. *Brussel*; *Nederlandsch*
mer van koophandel voor België en Luxen
Koningstr., 93; 1973, 4°, 287 *blz.* (*Gest*
630 *fr.*; *voor leden*, 490 *fr.*).
B.D. 3518 33 73

Catalogue Indexes

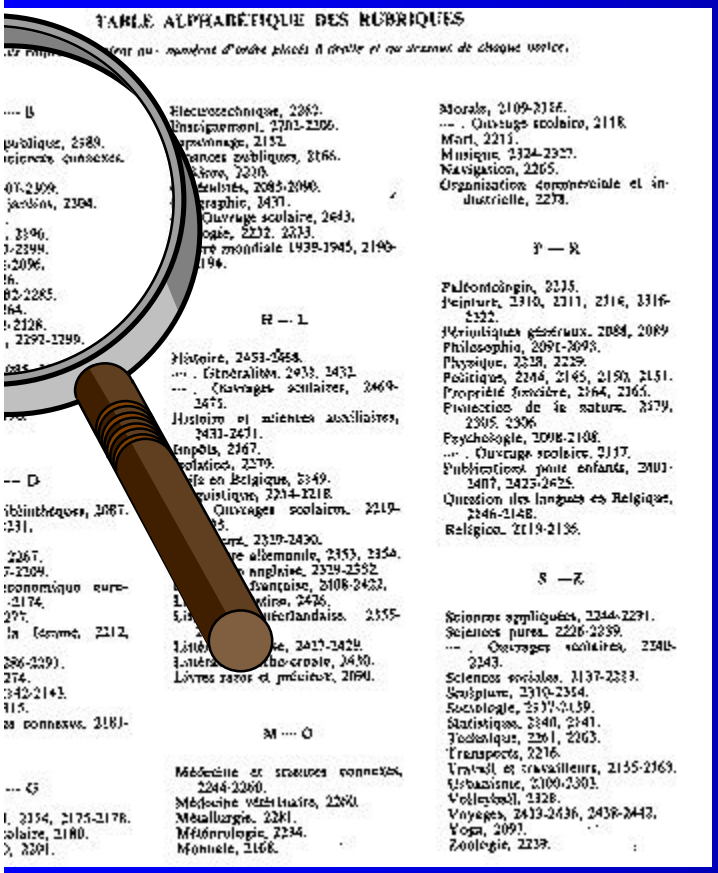


TABLE ALPHABÉTIQUE DES RUBRIQUES

Électrotechnique, 2262.
Enseignement, 2202-2206.
Morale, 2109-2116.
--- . Ouvrage scolaire, 2118.
Mort, 2211.
Musique, 2324-2327.
Navigation, 2265.
Organisation économique et industrielle, 2278.

P — R

Paléontologie, 2235.
Peinture, 2310, 2311, 2316, 2316-2322.
Périodiques généraux, 2088, 2089
Philosophie, 2091-2093.
Physique, 2228, 2229.
Politique, 2144, 2145, 2150, 2151.
Propriété foncière, 2264, 2265.
Protection de la nature, 2279, 2305, 2306
Psychologie, 2098-2108.
--- . Ouvrage scolaire, 2137.
Publications pour enfants, 2401-2407, 2425-2625.
Question des langues en Belgique, 2146-2148.
Religion, 2119-2136.

S — Z

Sciences appliquées, 2244-2291.
Sciences pures, 2226-2239.
--- . Ouvrage scolaires, 2240-2243.
Sciences sociales, 2137-2223.
Sculpture, 2319-2324.
Sociologie, 2177-2139.
Statistique, 2240, 2241.
Technique, 2261, 2263.
Transports, 2216.
Travail et travailleurs, 2155-2169.
Urbanisme, 2300-2303.
Volleyball, 2328.
Voyages, 2433-2436, 2438-2442.
Yoga, 2097.
Zoologie, 2239.

Administration publique, 2189
Agriculture et sciences complexes 2266, 2267
Architecture, 2307-2309
Architecture de jarsdins, 2304
Arpentage, 2227
ASSistance, 2195, 2196 ...

Record structures

optional

Top Left Corner

optional

Top Right Corner

Headers:

- Main area: Title and statement of Authorship - Edition - Imprint - Collation
- Supplementary area: series - Notes - ISBN Binding - Price - Number of copies

optional

Corner

optional

Bottom Right Corner

Danish Library

UDC

ODY

uthor / Title

ddress (location, publisher, year)

ollation(Material description of the work)

ollection(series, volume,...)

optional

OTE (about the title)

EFERENCE

ORDER

Belgian Library

VEDETTE

BODY

COTES (required only if each sub-reference of volume contains a cote group

Notes

State of the collection

optional

French Library

Application on Belgian records

159.962
ir (Gérard). Je suis fakir. ([Par] Gé-
-Belair). (Verviers, Editions Gérard &
32° carré, couv., ill., 158 p. (30 fr.).
-flash, 352).
ɔductif : Souvenirs, révélations, con-
4 352 73-2108

Pre-ISBD
~ 1973

UDC Heading	**<675 I=bb <$a 159.962</$a></675>**
Author	**<200 I=0b <$f Gérard Liger-Belair</ <$a Je suis fakir </$a></20** **<700 I=b0 <$a Liger-Belair</$a> <$b Gérard <$b></700>**
Publisher	**<210 I=bb <$a Verviers</$a> <$c Editions Gérard & C° <$d [1973]</$d></210>**
Collation	**<215 I=bb <$d 320 carré</$d> <$c couv., ill.</$c> <$a 158 p.</$a></215>**
Price	**<010 I=bb <$d 30 BEF</$d></010>**
Collection	**<225 I=2b <$a Marabout-flash <$v 352</$v></225>**
Note	**<517 I=0i1 <$a Souvenirs, révélations conseils</$a></517>**
Ref.	**<900 I=bb <$a B.D. 14.814352</$a> <$b 73-2108</$b></900>**

Extension to bibliographic references

] J. Juan, Y. Y. Tang, and C. Y. Suen. Four Directional Adjacency Graphs (fdag) a
ation in Locating Fields in Forms. In Third International Conference on Document
ognition (ICDAR'95), pages 752-755. IEEE Computer Society Press, Aug. 1995.

: J. Juan, Y. Y. Tang, and C. Y. Suen
: Four Directional Adjacency Graphs (fdag) and their Application in Locating fie
:
: Aug
: 1995
:
:
: IEEE Computer Society Press
:
: 752-755
on:
: Third International Conference on Document Analysis and Recognition (ICDA
:
: